Appeared in the 2025 International Joint Conference on Neural Networks (IJCNN 2025).

Federated Learning with Patient-Annotated Data in Epileptic Seizure Detection

Amin Aminifar[1,2], Jonathan Dan[1], David Atienza[1]
[1]Embedded Systems Laboratory, EPFL, Switzerland
[2]Institute of Computer Engineering, Heidelberg University, Germany
{amin.aminifar, jonathan.dan, david.atienza}@epfl.ch

***Abstract*—Machine learning (ML) generally requires a substantial amount of data to reach or surpass human-level performance. However, data collection and annotation by experts are known to be costly and time-consuming, which often leads to suboptimal performance for ML algorithms. One approach to tackle this challenge is to adopt patient-annotated data on each patient's device in a federated learning (FL) setting. However, this approach comes with certain challenges. For instance, in the case of epilepsy monitoring, patient-annotated data is known to involve inaccuracies, i.e., patients may lose consciousness and annotate a seizure with substantial delay compared to the seizure onset. To address this challenge, we propose an FL framework for epileptic seizure detection with noisy patient-annotated data. We evaluate our approach in the case of epileptic seizure detection and show that our proposed method achieves up to 32.63% higher accuracy, 32.95% higher specificity, and 22.28% higher F1 score compared to the model trained on the noisy dataset.**

***Index Terms*—federated learning, wearable devices, noise, electroencephalogram (EEG), seizure detection.**

I. Introduction

The performance of ML models is influenced by several factors, including the learning algorithm, the complexity of the model (e.g., the depth of deep neural networks) and design parameters such as the number of hidden nodes [1]. However, the most critical factor often lies in the data itself, as algorithms derive their models from this data. Key aspects such as the size and quality of the training data significantly affect the performance of the model [2]–[5]. Despite its importance, collecting large amounts of high-quality data poses significant challenges, particularly in the medical domain.

For medical data, one major challenge is the cost and effort required to collect and annotate accurately. For example, acquiring clean physiological signals such as electrocardiogram (ECG) or electroencephalogram (EEG) often requires the hospitalization of patients for extended periods. Expert clinicians must meticulously review and label the collected data. This process becomes even more challenging in conditions like epilepsy, where seizures are rare and unpredictable events. To capture data during seizures, patients may need to remain hospitalized for prolonged periods, often with reduced medication to increase the likelihood of occurrence of seizures. Moreover, obtaining sufficient data from multiple seizure episodes to enhance ML outcomes requires even longer hospitalization. These factors exacerbate the issue of limited data, which restricts the complexity and overall performance of ML models.

An approach to addressing this challenge is the use of wearable devices that allow patients to collect and annotate their data without requiring daily hospitalization. These systems enable patients to continue their routines at home, work, school, and other environments while collecting data continuously. Another advantage is that data collected in hospitals may not accurately reflect typical physiological patterns of a patient, since it only captures data while the patient is resting in a controlled hospital setting. In contrast, wearable systems provide data that more realistically represent the daily life of a patient. However, wearable systems only have access to the patient's data by using the device, limiting the data for accurate detection of seizures. To enable training of robust ML models, while protecting patient privacy and avoiding the inefficiency of transferring large volumes of data, FL frameworks have been adopted in several studies [6]–[8].

In FL on wearable systems for seizure detection, the data is typically annotated by patients rather than by medical experts. However, patient annotations are often highly unreliable. For example, in the case of epilepsy, seizures can be annotated late or completely missed if the patient is unconscious during the event [9]. This presents a significant challenge for the training of ML models. While methods for handling noisy labels have been explored in other applications [10]–[13], their application in the context of epilepsy and FL remains unexplored.

In this paper, we propose an FL framework designed for patient-annotated data in epilepsy. Our approach addresses the challenge of noisy labels, where patients can identify a general period during which a seizure occurred but may not precisely pinpoint the seizure event. To address this, we integrate well-established domain knowledge in ictal EEG signals into the FL process to estimate the confidence level for each label. This approach is inspired by [14], which highlights how the spectral power of EEG signals across various frequency bands can quantify changes associated with epileptic seizures, providing valuable insights for seizure detection. By leveraging these spectral analyses, we aim to refine seizure event identification within patient-annotated datasets. We evaluate our methodology using the publicly available Physionet CHB-MIT scalp EEG database [15], [16]. Our experimental results

This work was supported in part by the Swiss NSF, grant no. 10.002.812: "Edge-Companions: Hardware/Software Co-Optimization Toward Energy-Minimal Health Monitoring at the Edge," and in part by the EU's Horizon 2020 grant agreement no. 101017915 (DIGIPREDICT).

demonstrated that, for sample-based evaluation (described in Section IV-A5), our proposed method achieves up to 32.63% higher accuracy, 32.95% higher specificity, and 22.28% higher F1 score compared to the model trained on the noisy dataset. For event-based evaluation (described in Section IV-A5), our proposed method achieves up to 39.7% higher precision, 31.2% higher F1 score, and 173.0 fewer FP/24h compared to the model trained on the noisy dataset.

In summary, our work advances the state-of-the-art in addressing the noisy label problem in seizure detection in the context of wearable devices and FL. Our approach is scalable, practical, and easy to implement and requires minimal hyperparameter tuning. It offers a promising solution for mitigating the effects of label noise in real-world applications and can be seamlessly integrated into future FL frameworks for seizure detection.

II. Related Works

FL is a setting for training ML models in scenarios where training data is decentralized [17]. The term *federated learning* was introduced in 2016 by McMahan et al. [18], [19]. Unlike traditional centralized approaches, FL allows data to remain distributed across various clients or parties, such as patients' personal devices, while collaboratively learning a global model. This process is typically orchestrated by a central server, but ensures that raw data never leaves local devices. By adhering to the data minimization principle outlined in the data protection guidelines [20], [21], FL significantly mitigates the privacy risks inherent in centralized ML systems [19].

In recent years, FL has gained significant attention, particularly in privacy-sensitive domains such as healthcare, where data sharing is often restricted by ethical and regulatory requirements [22]–[24]. Although FL methods are commonly based on neural networks, the setting has been successfully extended to other ML algorithms, including tree-based approaches [25]–[29]. Comprehensive reviews of FL applications in healthcare are available in [30]–[33]. A notable recent application of FL in healthcare is its use in seizure detection [6]–[8], which is the main focus of this paper.

Epilepsy is a neurological disorder that affects around 50 million people worldwide, according to the World Health Organization (WHO) [34]–[36]. Despite advances in antiepileptic drugs, approximately one-third of people with epilepsy (PWE) continue to experience recurrent seizures. These seizures significantly affect their quality of life and, in severe cases, may result in sudden unexpected death in epilepsy (SUDEP) [37]. On average, PWE face a two- to three-fold higher risk of premature death compared to the general population [38]. Mobile health monitoring using wearable devices offers a promising solution for real-time seizure detection, allowing alerts to family members and caregivers for timely intervention [39]–[46]. Research on wearable devices for the monitoring of ambulatory epilepsy has progressed, with prototypes already being developed [14], [47].

The growing use of ambulatory and long-term EEG monitoring underscores the urgent need for high-quality automated seizure detection algorithms based on electroencephalography. Advances in ML and the availability of EEG datasets from PWE have driven progress in this field. However, annotated EEG datasets for seizures remain scarce and are often inaccessible due to strict legal requirements surrounding personal health data [48]. FL offers a promising approach to leverage seizure data collected by wearable systems while preserving data privacy. Yet, FL faces unique challenges as patient-annotated data is inherently noisy and unreliable. For example, seizures may be annotated late or completely missed, particularly if the patient is unconscious during the event [9]. Addressing these challenges is critical for advancing the reliability and effectiveness of FL in seizure detection.

The handling of noisy labels has been studied in [10]–[13]. For instance, [13] introduces Pi-DUAL, an innovative architecture designed to address label noise in deep learning by leveraging privileged information (PI) available exclusively during training. Pi-DUAL employs a dual-path network architecture consisting of a prediction network, optimized for clean labels, and a noise network, which focuses on handling noisy annotations. The gating mechanism in Pi-DUAL adaptively balances the influence of the clean prediction network and the noise network, minimizing the impact of noisy labels while leveraging privileged information. This design helps to improve model robustness and generalization. Although Pi-DUAL has proven effective in addressing label noise in general deep learning applications, its applicability to the specific challenges of seizure detection, particularly in FL settings with patient-annotated data, has not been explicitly explored.

The problem of annotation of seizures has been explored in [49], where a self-learning methodology for the detection of epileptic seizures without medical supervision is presented. The authors propose a minimally-supervised algorithm to label seizures on edge devices automatically. In their approach, patients confirm that a seizure occurred within the last hour by pressing a button on a smartphone or mobile device after recovering from the seizure. Additionally, the average duration of the patient's seizures, typically provided by medical experts, is incorporated into the labeling process. EEG signals are analyzed to extract features indicative of seizures, such as frequency-domain power features. Using these features, the position of the seizure within the EEG recording is determined through a clustering scheme. The identified seizure windows are then labeled based on the clustering results and the patient-reported average seizure duration. This method allows for a degree of automation in seizure annotation while still utilizing minimal patient input.

In [50], the authors introduce the Maximum-Mean-Discrepancy Decoder (M2D2), a method designed to help medical professionals by automatically localizing and labeling seizures in long EEG recordings. The approach processes a lengthy EEG signal and identifies a timestamp t, indicating that a seizure probably occurred within $t \pm \Delta$ minutes. This significantly reduces the search effort for experts, who only need to review a 2Δ-minute interval instead of the entire recording. Unlike traditional methods that rely on patient-

specific data or manual labels, M2D2 employs a statistical tool called maximum-mean-discrepancy (MMD). This technique enables the system to detect regions prone to seizures by comparing patterns in the EEG signal, identifying segments most likely associated with seizures. This approach streamlines the annotation process and reduces the dependency on manual intervention.

In summary, the advancements in FL, seizure detection, and methods for handling noisy labels have laid a strong foundation for addressing critical challenges in healthcare applications. However, significant gaps remain, particularly in managing noisy patient-annotated data within FL frameworks for seizure detection. By addressing these challenges, our work aims to enhance the reliability and effectiveness of FL for seizure detection, paving the way for scalable, real-world implementations that improve patient outcomes.

III. Method

In this section, we present the details of our approach. We begin with a brief motivation and formalization of the problem, followed by a detailed explanation of our proposed methodology.

A. Motivation and Problem Formulation

Currently, the gold standard for epilepsy monitoring is video-electroencephalogram, which combines EEG recordings of brain activity with closed-circuit video observation. However, long-term EEG monitoring outside hospital settings is challenging. Scalp-EEG systems, such as hats and caps, are intrusive and can cause discomfort and social stigmatization, limiting their suitability for extended monitoring periods [51]. Although intracerebral or subcutaneous EEG allows for very long-term monitoring, it is invasive, expensive, and viable for only a small subset of patients [52], [53].

Recent advances have introduced lightweight, non-stigmatizing wearable systems, with prototypes already available [14], [47]. These systems not only enable real-time patient monitoring, but also facilitate the collection of data during daily life activities. This approach has dual benefits: it allows the collection of epilepsy data on a broader scale and provides a more accurate representation of patients' real-life conditions, compared to data collected in controlled environments. However, utilizing data collected from such wearable systems for training ML models to detect epileptic seizures presents several challenges. First, the data is inherently decentralized and cannot be centralized due to regulatory privacy requirements. FL frameworks are thus necessary to analyze this distributed data while maintaining privacy and complying with regulations. Second, the data is annotated by patients rather than medical experts, leading to inaccuracies in labels. Patients may not precisely identify seizure events, either due to delays in annotation or inability to pinpoint exact timestamps, further complicating the training of ML models.

Problem Formulation: We consider an FL setting involving n clients, where each client represents a patient. Patients collect their EEG signals using wearable devices, such as the e-Glass system [14], which utilizes four electrodes to record two-channel EEG signals. The overall FL scenario is illustrated in Figure 1.

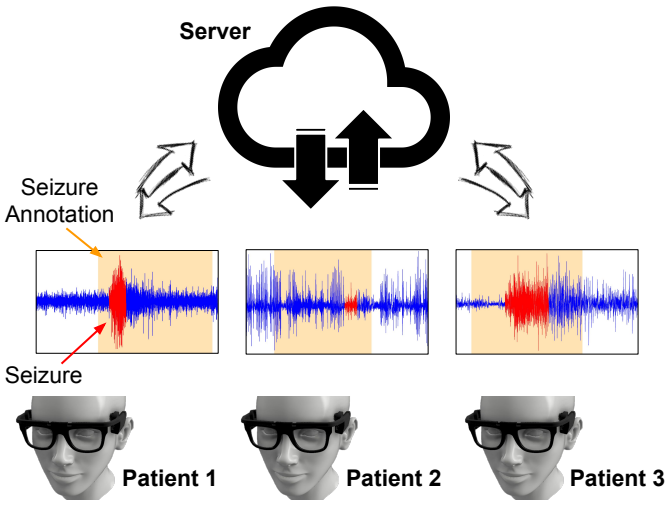


Fig. 1: Overview of the Scenario

In this setting, when no seizure event occurs, the patient does not report a seizure, allowing negative samples to be correctly labeled. However, during a seizure event, patients can only approximate the seizure timestamps. This limitation arises because patients often lack normal consciousness during seizures. We assume that the patient annotates a duration of D minutes (e.g., 10 minutes) within which the seizure is believed to have occurred, without providing precise timestamps.

For patient-reported labels, negative samples are assumed to be accurate and used as-is. Conversely, positive labels are considered unreliable but serve as a starting point for deriving more accurate annotations.

B. Frequency Band Analysis and Epilepsy

Normal human EEG activity typically ranges from 1–30 Hz in frequency, with amplitudes between 20–100 μV. This range is divided into distinct frequency bands: alpha (8–13 Hz), beta (13–30 Hz), theta (4–7 Hz), and delta (0.5–4 Hz). The surface EEG shows patterns that correspond to various states such as sleep, wakefulness, and specific pathophysiological processes like seizures. These patterns are defined by the frequency and amplitude of the electrical signals [54].

For example, alpha waves of moderate amplitude are associated with relaxed wakefulness and are most prominent in the parietal and occipital regions. In contrast, beta activity with lower amplitude is more prominent in the frontal areas and increases during intense mental activity. When a relaxed individual becomes alert, the EEG undergoes desynchronization, marked by a decrease in alpha waves and an increase in beta activity. Theta and delta waves are typically seen during drowsiness and early stages of slow-wave sleep, but their presence during wakefulness can indicate brain dysfunction [54].

The relationship between EEG signal power in specific frequency bands and epileptic seizures has been explored in studies such as [14], [49], [50]. In particular, an increase in power within the delta and theta frequency bands is often associated with seizure events. During seizure episodes, the normalized power of the delta band shows a marked increase, demonstrating a strong correlation between frequency band power, particularly in the delta band, and seizure activity.

C. Frequency Band Power Calculation

To estimate the Power Spectral Density (PSD) of EEG signals, we can use a method such as the Welch method [55]. In this method, the signal is divided into overlapping segments and windowed using a function such as the Hanning window [56]. The PSD is computed for each segment, and the average is taken:

$$P(f_n) = \frac{1}{K}\sum_{k=1}^{K} I_k(f_n),$$
$$I_k(f_n) = \frac{1}{LU}\left|\sum_{j=0}^{L-1} X_k(j)W(j)e^{\frac{-2\pi ijn}{L}}\right|^2,$$

here, $X_k(j)$ represents the k-th segment of the signal, $W(j)$ is the window function, and U is the normalization factor for the window (defined as: $U = \frac{1}{L}\sum_{j=0}^{L-1} W(j)^2$). Band power for specific frequency bands (e.g., delta: $0.5\,\text{Hz}$–$4\,\text{Hz}$) is computed by integrating the PSD over the desired frequency band:

$$\text{Band Power} = \int_{f_{\text{low}}}^{f_{\text{high}}} P(f)\,df.$$

D. Frequency Band Power As Metric for Recognizing Noise

As described in Section III-A, positive annotations are noisy, and this frequency band power metrics are used to address them. For each patient and each seizure event, D minutes of data are annotated, within which the seizure is known to have occurred. The EEG signals within this D-minute duration are divided into several windows.

For each window, we calculate the delta band power (similarly, this process can be applied to the theta frequency band). The values for the windows within D are then transformed as follows: first, we compute the mean (μ) and standard deviation (σ) of the values. Next, we standardize the values using the formula $z = \frac{x-\mu}{\sigma}$. After standardization, we rescale the values (originally centered around zero) using $\frac{z+1}{2}$. Finally, the rescaled values are clipped to ensure they fall within the range zero and one. We denote this transformed metric as $\mathcal{M}^i_{fb}$, where fb represents the frequency band, and i indicates that the metric corresponds to the i-th window within D.

The obtained values can be interpreted as confidence scores for the positive labels of the windows. This metric can be calculated in an FL setting, as each patient only needs access to their local data for computation. Based on these calculated metrics, several approaches can be taken to address noisy labels. A straightforward approach involves calculating the mean (μ') and standard deviation (σ') of the metrics (after standardization and transformation) for all windows within D. Then a threshold τ can be defined, for example, $\tau = \mu' + \sigma'$, to classify windows, similar to [44], [57]. If $\mathcal{M}^i_{fb} > \tau$, the window and its label can be used in the training process as is. Conversely, if the metric is below the threshold, the confidence in the label decreases. In such cases, the sample can either be removed, or its label flipped to zero, and the modified samples can still be used during training.

E. Federated Learning

1) Federated Training Procedure: We employ the federated stochastic gradient descent (FedSGD) procedure introduced in [18], [19] to train the deep learning model. In this framework, we have multiple clients, representing patient devices that hold their local data, and a central server responsible for coordinating the training process.

Our procedure involves the following steps:

(1) *Band Analysis for Noisy Labels:* Each client performs a frequency band analysis on their local dataset to address labels' noise. The results of this analysis are used to clean the dataset by either removing noisy samples or flipping the labels of these samples. The cleaning method is determined by an agreed-upon protocol established before training begins. The cleaned dataset for client C_i is denoted as $\mathcal{D}^{C_i}$.

(2) *Model Download:* Each client downloads the current global model w from the server.

(3) *Mini-Batch Selection:* Clients select a mini-batch of size s (predefined at the start of training) from their cleaned dataset for the current round j. For client C_i, this mini-batch is denoted as $b_j^{C_i}$.

(4) *Forward Pass:* Each client performs a forward pass using their mini-batch $b_j^{C_i}$ and the global model w to compute predictions.

(5) *Gradient Calculation:* Clients calculate gradients based on a shared loss function ℓ (e.g., cross-entropy loss), which is also agreed upon before training begins. The gradient for client C_i in round j is denoted as: $\nabla_j^{C_i} = \nabla\ell(w, b_j^{C_i})$.

(6) *Gradient Scaling:* Each client scales its calculated gradient by the size of its cleaned dataset $|\mathcal{D}^{C_i}|$: $\nabla^{C_i}_{\text{scaled},j} = \nabla_j^{C_i} \cdot |\mathcal{D}^{C_i}|$. The scaled gradient is then sent to the server.

(7) *Server Aggregation:* The server collects scaled gradients from all K clients and computes the aggregated gradient by dividing the sum of these gradients by the total number of samples $|\mathcal{D}|$, where $\mathcal{D} = \sum_{i=1}^{K}|\mathcal{D}^{C_i}|$. The aggregated gradient is computed as: $\nabla_j = \frac{1}{|\mathcal{D}|}\cdot\sum_{i=1}^{K}\nabla^{C_i}_{\text{scaled},j}$.

(8) *Model Update:* The server updates the global model w using an optimization algorithm with a learning rate η: $w \leftarrow w + \eta\nabla_j$.

(9) *Model Distribution:* The updated global model is distributed to all clients for the next round. This process repeats starting from Step 2, continuing either for a predefined number of rounds or until a convergence criterion is met (e.g., validation set performance).

This iterative process enables collaborative training of the global model across all clients while maintaining data privacy, as raw data remains on the clients' devices. To further enhance

privacy and prevent the sharing of raw gradients, we incorporate a secure multiparty computation scheme within our FL framework.

2) Weighted Loss: Cross-entropy loss is a common choice for classification tasks as it measures the difference between the predicted probability distribution and the true class labels. Seizure datasets are often substantially imbalanced; in such cases, using standard cross-entropy loss, which treats all classes equally, may lead to higher specificity but reduced sensitivity for the minority class. To address this, weighted cross-entropy can be employed to assign greater importance to the positive class during training. However, in seizure detection, maintaining high specificity (low false positive rate) is critical. The use of a weighted loss function should be guided by the characteristics of the dataset. For instance, if the dataset does not exhibit a severe imbalance between negative and positive samples, applying more weight to the negative samples can help prioritize specificity while balancing sensitivity.

Weighted cross-entropy loss mitigates class imbalance by assigning weights to classes, ensuring a greater contribution of a desired class during training. The formula is: $\ell(x, y) = \frac{1}{N}\sum_{n=1}^{N} -w_{y_n} \log \frac{\exp(x_{n,y_n})}{\sum_{c=1}^{C}\exp(x_{n,c})}$, where N is the batch size, C the number of classes, $x_{n,c}$ the model output (logit) for the n-th sample and c-th class, y_n the true label and w_{y_n} the weight of the class, typically calculated as $w_c = 1/f_c$, with f_c being the frequency of class c.

IV. Evaluation

A. Experimental Setup

1) Dataset: For our evaluation, we utilize the publicly accessible Physionet CHB-MIT scalp EEG database [15], [16]. This dataset comprises EEG recordings from pediatric patients with intractable seizures. These patients were monitored over several days following the withdrawal of anti-seizure medications to document their seizures and evaluate their eligibility for surgical intervention. The dataset includes recordings grouped into 24 cases, collected from 23 individuals. In this study, we focus on a wearable healthcare context, such as the e-Glass system [14], and therefore only consider the four electrodes from the $F7-T7$ and $F8-T8$ channels. Consequently, three files from case chb12, which contain unipolar recordings, are excluded. The resulting dataset includes 185 seizure events across all patients [58], [59].

Choice of benchmark dataset and EEG channels: Despite CHB-MIT's limited size, this dataset remains widely adopted in seizure detection studies due to its accessibility and the availability of clinically annotated seizure events [8], [24], [42], [50], [60], [61]. In this research, we only considered four electrodes and two channels, i.e., $F7$-$T7$ and $F8$-$T8$ channels. This choice reflects the focus of our study on wearable settings, where the number of electrodes is typically limited. This limitation stems from the need for wearable devices to be compact, unobtrusive, and suitable for ambulatory applications. One such device in the context of seizure detection is the e-Glass system [14]. Specifically, the e-Glass system uses electrodes placed at positions corresponding to the $F7$-$T7$ (between $F7$ and $T7$/$T3$) and $F8$-$T8$ (between $F8$ and $T8$/$T4$) channels in the CHB-MIT database.

2) Data Preparation: We make a noisy dataset based on the problem we described in III-A, with the specifics outlined here. Each patient's data consists of multiple files, some containing seizure events and others that do not. For each seizure event, we extract D minutes of EEG signals surrounding the event, labeling this segment with a noisy positive annotation. The exact position of the seizure event within this D-minute window is random.[1] After selecting segments with noisy positive annotations, we randomly extract non-seizure periods of equal duration from files without seizure events, assigning them clean negative annotations. In this dataset, we utilize 4-second windows, consistent with previous research [14], [61], with a 2-second (50%) overlap between consecutive windows.

We adopt the leave-one-out cross-validation (LOOCV) procedure for training and testing.[2] In each experiment, one case from chb01 to chb24 is designated for testing, while the remaining cases are used for training. We use our generated data set with noisy annotations to train the deep learning models. However, for testing and to allow for a comprehensive evaluation, we use all recording files rather than restricting to $2 \cdot D$ minutes per seizure. This approach leads to a highly imbalanced test set, significantly affecting some evaluation metrics. For example, the disproportionate number of negative samples may notably reduce the F1 score.

Type of label noise: Our method targets a specific type of label noise, namely, temporal imprecision in positive event annotation caused by patients losing consciousness or responding with delay. By leveraging frequency band power analysis, we infer a confidence level for each positive label. This mechanism improves model performance even when a large proportion of the training data contains noisy labels. While this work focuses on noise due to annotation delay, our framework could potentially be adapted to handle other types of label noise, such as mislabeling or random corruption, by adjusting the confidence metric or incorporating additional strategies.

3) Deep Neural Network Model: For our neural network architecture, we employ the fully convolutional network (FCN) introduced in [61]. The base of the FCN consists of three sequential blocks, each comprising a convolutional layer, a batch normalization layer, rectified linear unit (ReLU) activation functions, and a pooling layer. Following these three blocks, the network incorporates two fully convolutional layers, and a softmax output layer.

[1] Since the start and end of the D-minute window are selected randomly, if the seizure event occurs near the beginning/end of a file, the extracted period may be shorter than D. Furthermore, if two seizure events occur close to each other within the same file, the D-minute windows may overlap. For seizure events longer than D, only D minutes are considered; however, such cases were rare in our experiments.

[2] In our evaluation, only windows that entirely fall within the ground truth seizure annotations of the CHB-MIT dataset are considered positive for correct annotation.

TABLE I: Sample-Based Evaluation (LOOCV): B1:Clean Data, B2:Noisy Data, M1:Noisy Data (our method: flip labels), M2:Noisy Data (our method: remove samples/weighted loss)

Patient	Accuracy				Specificity				Sensitivity				F1 score			
	B1	B2	M1	M2	B1	B2	M1	M2	B1	B2	M1	M2	B1	B2	M1	M2
chb01	99.87%	60.02%	99.28%	99.85%	99.96%	59.91%	99.34%	99.98%	67.29%	98.13%	76.64%	55.14%	74.61%	1.42%	38.32%	68.21%
chb02	99.67%	56.89%	98.9%	99.45%	99.7%	56.83%	98.92%	99.48%	80.49%	100.0%	82.93%	73.17%	38.82%	0.6%	16.35%	25.53%
chb03	99.61%	66.83%	97.81%	99.4%	99.69%	66.82%	98.04%	99.66%	71.88%	71.88%	16.67%	8.33%	51.11%	1.2%	4.1%	7.27%
chb04	96.31%	72.09%	95.74%	96.65%	96.34%	72.07%	95.78%	96.69%	41.3%	90.22%	44.57%	40.22%	1.44%	0.42%	1.35%	1.55%
chb05	99.81%	74.02%	99.14%	99.75%	99.98%	73.94%	99.26%	99.95%	55.15%	94.85%	68.38%	46.32%	69.12%	2.75%	38.04%	58.6%
chb06	99.87%	44.44%	99.4%	99.82%	99.92%	44.46%	99.45%	99.87%	0.0%	13.33%	0.0%	0.0%	0.0%	0.02%	0.0%	0.0%
chb07	98.28%	82.76%	99.21%	99.48%	98.32%	82.74%	99.25%	99.53%	74.68%	92.41%	69.62%	60.76%	10.23%	1.38%	18.84%	23.36%
chb08	99.08%	78.37%	99.28%	99.28%	100.0%	78.17%	99.83%	99.96%	26.99%	94.25%	55.75%	45.58%	42.51%	9.86%	65.97%	61.31%
chb09	99.58%	83.36%	99.3%	99.49%	99.59%	83.34%	99.3%	99.49%	96.97%	100.0%	96.97%	95.45%	33.51%	1.28%	22.98%	28.77%
chb10	99.94%	55.03%	99.54%	99.81%	100.0%	55.09%	99.64%	99.98%	74.29%	26.67%	58.1%	24.76%	84.32%	0.28%	37.31%	37.68%
chb11	99.71%	76.35%	99.43%	99.46%	100.0%	76.23%	99.68%	99.82%	54.27%	95.98%	60.3%	44.22%	70.13%	4.91%	57.14%	51.16%
chb12	98.73%	38.28%	95.97%	98.17%	99.8%	37.73%	96.99%	99.25%	10.71%	83.48%	12.5%	9.38%	16.9%	3.15%	6.95%	10.97%
chb13	98.61%	54.16%	95.37%	98.46%	98.98%	54.14%	95.7%	98.82%	10.4%	60.0%	19.2%	12.0%	5.92%	1.09%	3.38%	6.15%
chb14	99.84%	69.59%	99.74%	99.73%	99.99%	69.67%	99.89%	99.88%	0.0%	16.67%	0.0%	0.0%	0.0%	0.17%	0.0%	0.0%
chb15	98.42%	62.89%	96.72%	97.89%	99.76%	62.81%	97.99%	99.2%	0.62%	69.14%	4.12%	1.85%	1.05%	4.79%	3.28%	2.31%
chb16	96.98%	62.06%	96.05%	97.42%	97.06%	62.06%	96.13%	97.5%	0.0%	71.43%	0.0%	0.0%	0.0%	0.31%	0.0%	0.0%
chb17	98.94%	69.28%	96.9%	98.6%	99.25%	69.2%	97.12%	98.9%	16.67%	90.28%	37.5%	22.22%	10.67%	2.19%	8.42%	10.81%
chb18	99.62%	77.18%	98.94%	99.76%	99.72%	77.14%	99.06%	99.92%	55.41%	93.24%	48.65%	31.08%	40.2%	1.85%	17.52%	37.4%
chb19	99.83%	77.41%	99.61%	99.81%	99.92%	77.42%	99.68%	99.95%	55.17%	72.41%	65.52%	37.93%	57.66%	1.36%	41.76%	46.81%
chb20	99.67%	60.29%	97.69%	99.4%	99.9%	60.43%	97.86%	99.65%	16.42%	10.45%	35.82%	8.96%	21.36%	0.14%	7.72%	7.5%
chb21	99.54%	85.47%	97.77%	99.34%	99.68%	85.47%	97.9%	99.5%	10.42%	83.33%	16.67%	0.0%	6.8%	1.83%	2.37%	0.0%
chb22	99.93%	67.31%	98.94%	99.71%	99.96%	67.27%	98.98%	99.75%	83.67%	87.76%	73.47%	77.55%	80.39%	0.93%	19.51%	48.1%
chb23	99.67%	63.72%	97.85%	98.54%	99.94%	63.72%	98.11%	98.85%	34.0%	64.0%	37.0%	26.0%	45.95%	1.45%	12.61%	13.0%
chb24	99.31%	56.61%	98.05%	98.12%	99.51%	56.4%	98.19%	98.25%	67.77%	88.43%	76.86%	76.86%	55.41%	2.51%	33.27%	34.0%
Average	99.2%	66.43%	98.19%	99.06%	99.46%	66.38%	98.42%	99.33%	41.86%	73.68%	44.05%	33.24%	34.09%	1.91%	19.05%	24.19%

4) Implementation Details: In our FL framework, each client represents an individual patient with access solely to their own data. Consequently, as described in IV-A2, given the 24 cases in the dataset and our use of the LOOCV procedure, 23 clients participate in each experiment. The training process is governed by a predefined stopping criterion of 100 iterations, meaning that the model undergoes 100 updates per experiment. We implemented our FL framework using the `PyTorch` library. The cross-entropy loss function was employed for training, with the Adam optimizer applied for parameter updates using its default settings. For frequency band analysis metrics, we utilized `scipy.signal.welch` to compute the power spectral density. Numerical integration was performed using the trapezoidal rule implemented through `numpy.trapz`.

5) Evaluation Metrics: For the evaluation, we adopt both sample-based and event-based approaches. Sample-based evaluation has traditionally been the primary method for assessing ML algorithms and seizure detection models. While it provides detailed insights into performance at the level of individual data samples, it does not address clinically relevant questions. Event-based evaluation has emerged as a complementary approach, focusing on clinical considerations such as accurately counting seizures and minimizing false alarms. This method is more in line with the real world requirements of seizure detection in clinical and home monitoring systems [48].

Sample-Based Evaluation: Sample-based scoring calculates performance metrics on a sample-by-sample basis. It evaluates the performance of seizure detection models by comparing the predicted label for each test set window against the corresponding ground truth label. For sample-based evaluation, we report metrics including accuracy, specificity (true negative rate), sensitivity (recall or true positive rate), F1 score, and precision.

Event-Based Evaluation: Event-based scoring evaluates seizures as events, relying on the overlap between reference and hypothesis annotations. Any overlap is considered a correct detection (TP), while hypothesis events that do not overlap with a reference event are counted as false positives (FP). We apply a moving average filter to smooth the predictions before performing event-based evaluations. Consistent with [42], we use a filter size of 11 (corresponding to 20 seconds).[3] For the event-based evaluation, we report sensitivity, F1 score, precision, and FP/24 (false positive per day rate).[4]

B. Experimental Results

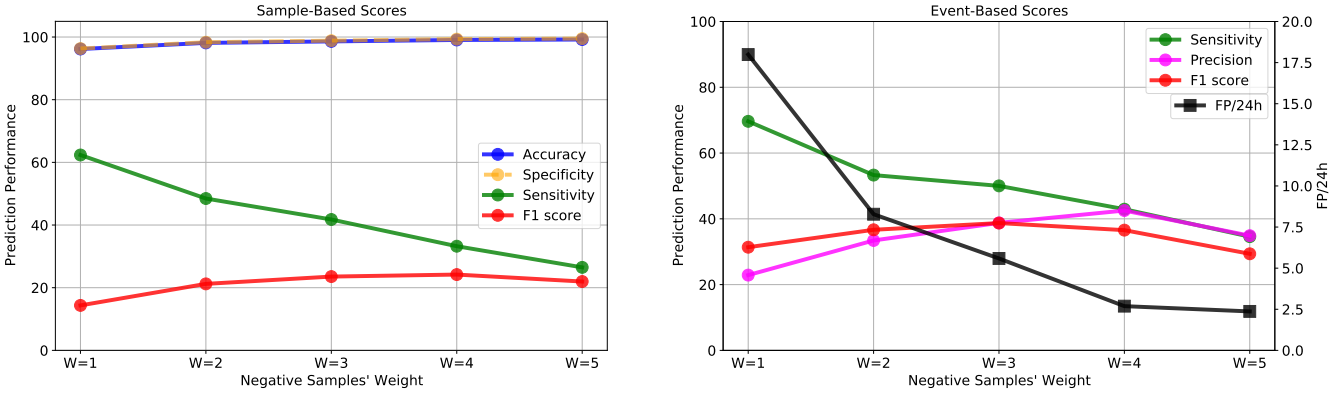
In this section, we present our experimental results based on the setup described in Section IV-A.

1) Performance Analysis: For the evaluation, we consider two main baselines and two variants of our proposed method, discussed below.

(1) *Clean Dataset:* In this baseline, we utilize the dataset described in Section IV-A2, but with clean labels for training. For this experiment, $D = 10$ min. The sample-based and event-based results are presented in Table I and Table II (B1 columns), respectively. This baseline is implemented based on FCN [61]. The authors achieved an F1 score of 46.6% on average, in a centralized scenario using all channels. In this work, we consider only two channels, aligning with the constraints of a wearable systems, in FL setting, which leads to an average F1 score of 34.09%.

[3]https://gitlab.epfl.ch/ashahbaz/personalized-online-seizure-detection

[4]For event-based scoring, we utilized the framework available at https://github.com/esl-epfl/timescoring.



(a) Sample-Based Scores

(b) Event-Based Scores

Fig. 2: The Effect of Adjusting the Weight Assigned to Negative Samples in the Cross-Entropy Loss

TABLE II: Averaged Results for Event-Based Evaluation (LOOCV): B1:Clean Data, B2:Noisy Data, M1:Noisy Data (our method: flip labels), M2:Noisy Data (our method: remove samples/weighted loss)

Metric	B1	B2	M1	M2
Sensitivity	50.1%	91.8%	51.6%	42.93%
Precision	46.57%	2.82%	36.37%	42.52%
F1 score	43.17%	5.36%	37.26%	36.56%
FP/24h	2.53	175.69	6.63	2.69

(2) *Noisy Dataset:* In this baseline, we utilize the dataset described in Section IV-A2, which includes noisy positive labels for training. For this experiment, $D = 10$ min. The sample-based and event-based results are presented in Table I and Table II (B2 columns), respectively.

(3) *Our Method (flip labels):* In this experiment, as outlined in III-D, we compute $\mathcal{M}^i_{fb}$ for all windows i within D (considering only samples annotated as positive by the patient). If $\mathcal{M}^i_{fb} > \tau$, where $\tau = \mu' + \sigma'$, the sample is retained with its original label; otherwise, its label is flipped from 1 to 0. For this experiment, $D = 10$ min and fb corresponds to the delta band. Using our band analysis approach, 88.9% of noisy samples and 58.3% of clean positive samples were correctly detected. The sample-based and event-based results are presented in Table I and Table II (M1 columns), respectively.

(4) *Our Method (remove samples):* In this experiment, as described in III-D, we compute $\mathcal{M}^i_{fb}$ for all windows i within D (considering only samples annotated as positive by the patient). If $\mathcal{M}^i_{fb} > \tau$, where $\tau = \mu' + \sigma'$, the sample is retained in the training set; otherwise, it is removed or ignored. For this experiment, $D = 10$ min and fb corresponds to the delta band. The sample-based and event-based results are presented in Table I and Table II (M2 columns), respectively.

In the sample-based evaluation, the results demonstrate that our proposed method (trained on noisy dataset), which involves removing samples during training (M2 column), achieves up to 32.63% higher accuracy, 32.95% higher specificity and 22.28% higher F1 score compared to the model trained on noisy dataset (B2 column). While the sensitivity is 40.44% lower, this is an expected trade-off between specificity and sensitivity. This trade-off is further influenced by the noisy dataset's composition, where 50% of the samples are labeled positive, the majority of which are noise. This skews the model, biasing it toward making more positive predictions. This is also true when comparing the model to the model trained on the clean dataset. When compared to the clean dataset (B1 column), our method exhibits only 0.14% lower accuracy, 0.13% lower specificity, 8.62% lower sensitivity, and 9.9% lower F1 score. These differences can be attributed to the significant noise present in the dataset, which impacts the overall performance.

In the event-based evaluation, our proposed method, which involves removing samples during training (M2 column), achieves up to 39.7% higher precision, 31.2% higher F1 score, and 173.0 fewer FP/24h compared to the model trained on the noisy dataset (B2 column). Comparable to the sample-based evaluation, the sensitivity is 48.87% lower. As explained earlier, this is due to the high proportion of noisy samples with positive labels, which biases the model toward making more positive predictions. Compared to the clean dataset (B1 column), our method shows only a 7.17% decrease in sensitivity, 4.05% lower precision, and 6.61% lower F1 score, with a marginal increase of 0.16 in FP/24h. These differences, as discussed previously, can be attributed to the substantial noise present in the dataset, which affects the overall performance.

ML performance in CHB-MIT: When performing LOOCV on the CHB-MIT dataset, similar to the evaluation setting in [61], from which we adopt the neural network architecture, we observed variability in model performance across cases (e.g., chb06, chb14). This observation aligns with prior research, which reports that seizure detection accuracy can vary significantly between patients due to multiple factors. These

discrepancies can be attributed to the limited seizure data per patient, variations in seizure morphologies, or seizure foci not properly captured by the selected EEG channels [62], [63]. The results reported in [61], also demonstrate such inter-patient variability and low performance for certain cases, even when a larger number of EEG channels are considered in the deep learning model.

2) Ablation Study: In this section, we analyze the effect of adjusting the weights assigned to positive and negative samples in the cross-entropy loss for our method, which involves sample removal. Since we are training on a subset of the dataset, as described in Section IV-A2, and removing certain samples, this approach might reduce the specificity. In the context of seizure detection, maintaining high specificity is crucial, as lower specificity translates to increased false alarms, which we aim to minimize. To address this, we increase the weight of negative samples during training to evaluate the impact of this adjustment on model performance and identify the most appropriate weight configuration for our application.

Figure 2 presents the results of adjusting the weights assigned to negative and positive samples in the cross-entropy loss. The evaluation starts with equal weights and progresses to assigning the negative samples a weight five times greater than that of the positive samples. Results are reported for both sample-based and event-based evaluations. The analysis shows that increasing the weight of negative samples improves metrics such as specificity and FA/24h, which correspond to a reduction in false alarms. However, this improvement comes at the cost of reduced sensitivity. As there is an inherent trade-off between sensitivity and specificity, selecting a weight configuration that meets the desired requirements is essential. In this study, we select $w = 4$ as it achieves a low FA/24h while remaining comparable to the model trained on clean data, as discussed in Section IV-B.

In this section, we also analyze the variation in D from 10 minutes to 40 minutes. Figure 3 presents the results based on the F1 score (event based) for the discussed baselines $B1$ and $B2$, as well as our proposed methods $M1$ and $M2$ (with different values of D). The results indicate that, as D increases, our approach continues to significantly outperform $B2$. However, the problem becomes more challenging with larger D, leading to a decline in the predictive performance of $M1$ and $M2$. Additionally, the decrease in $B1$'s performance from $D = 20$ minutes to $D = 40$ minutes could be attributed to changes in our dataset.

V. Conclusion

In this paper, we have addressed a critical challenge in FL for wearable seizure detection systems using patient-annotated data. We highlighted the significant noise present in patient-annotated data and proposed a method to formulate and address this issue effectively. Our experimental results, based on the publicly available CHB-MIT benchmark dataset, demonstrated that for sample-based evaluation, our proposed method achieves up to 32.63% higher accuracy, 32.95% higher specificity, and 22.28% higher F1 score compared to the model trained on the noisy dataset. For event-based evaluation, our proposed method achieves up to 39.7% higher precision, 31.2% higher F1 score, and 173.0 fewer FP/24h compared to the model trained on the noisy dataset.

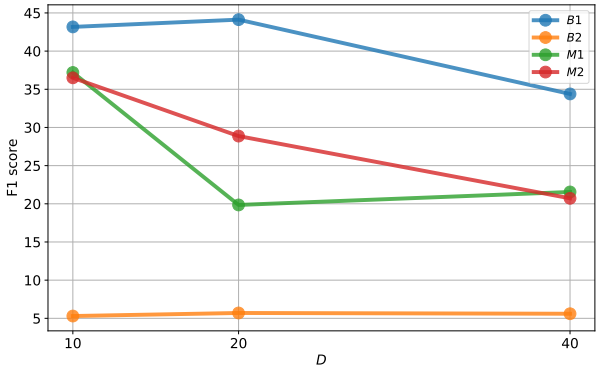


Fig. 3: Variation in D from 10 to 40 minutes

In this study, our focus was on EEG signals. The primary reason is that EEG is one of the most widely used biosignals for seizure detection, and recent advancements have made ambulatory EEG monitoring feasible using state-of-the-art wearable devices. Our approach to addressing label noise is inspired by neuroscience research, which shows that various physiological states, such as sleep, wakefulness, and patho-physiological processes like seizures, show distinct patterns in EEG signals. Although the current implementation relies on EEG-specific spectral features, our framework may be extended to support building models based on other biosignals.

This work advances the state-of-the-art in mitigating the noisy label problem for seizure detection in wearable devices within an FL context. Moreover, it provides a practical solution for managing label noise in real-world scenarios and can be easily integrated into future FL frameworks for seizure detection.

References

[1] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 215–223.

[2] V. Gudivada, A. Apon, and J. Ding, "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations," *International Journal on Advances in Software*, vol. 10, no. 1, pp. 1–20, 2017.

[3] A. Ng, "Machine learning yearning: Technical strategy for ai engineers, in the era of deep learning," 2018.

[4] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PloS one*, vol. 14, no. 11, p. e0224365, 2019.

[5] D. Rajput, W.-J. Wang, and C.-C. Chen, "Evaluation of a decided sample size in machine learning applications," *BMC bioinformatics*, vol. 24, no. 1, p. 48, 2023.

[6] S. Baghersalimi, T. Teijeiro, D. Atienza, and A. Aminifar, "Personalized real-time federated learning for epileptic seizure detection," *IEEE journal of biomedical and health informatics*, vol. 26, no. 2, pp. 898–909, 2021.

[7] S. Baghersalimi, T. Teijeiro, A. Aminifar, and D. Atienza, "Decentralized federated learning for epileptic seizures detection in low-power wearable systems," *IEEE Transactions on Mobile Computing*, 2023.

[8] A. Aminifar, M. Shokri, and A. Aminifar, "Privacy-preserving edge federated learning for intelligent mobile-health systems," *Future Generation Computer Systems*, vol. 161, p. 625–637, Dec. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.future.2024.07.035

[9] B. Blachut, C. Hoppe, R. Surges, C. Elger, and C. Helmstaedter, "Subjective seizure counts by epilepsy clinical drug trial participants are not reliable," *Epilepsy & Behavior*, vol. 67, pp. 122–127, 2017.

[10] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE transactions on neural networks and learning systems*, vol. 34, no. 11, pp. 8135–8153, 2022.

[11] M. Collier, R. Jenatton, E. Kokiopoulou, and J. Berent, "Transfer and marginalize: Explaining away label noise with privileged information," in *International Conference on Machine Learning*. PMLR, 2022, pp. 4219–4237.

[12] G. Ortiz-Jimenez, M. Collier, A. Nawalgaria, A. N. D'Amour, J. Berent, R. Jenatton, and E. Kokiopoulou, "When does privileged information explain away label noise?" in *International Conference on Machine Learning*. PMLR, 2023, pp. 26 646–26 669.

[13] K. Wang, G. Ortiz-Jimenez, R. Jenatton, M. Collier, E. Kokiopoulou, and P. Frossard, "Pi-DUAL: Using privileged information to distinguish clean from noisy labels," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235. PMLR, 2024, pp. 51 214–51 236.

[14] D. Sopic, A. Aminifar, and D. Atienza, "e-glass: A wearable system for real-time detection of epileptic seizures," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2018, pp. 1–5.

[15] A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[16] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[17] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[18] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[19] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1273–1282.

[20] W. House, "Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy," *White House, Washington, DC*, 2012.

[21] "Principles of the GDPR," https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr_en, 2024, [Accessed: Dec. 30, 2024].

[22] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 13, no. 3, pp. 1–207, 2019.

[23] S. Silva, B. A. Gutman, E. Romero, P. M. Thompson, A. Altmann, and M. Lorenzi, "Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 270–274.

[24] A. A. Fahliani, A. Aminifar, and A. Aminifar, "Privacy-preserving federated interpretability," in *IEEE International Conference on Big Data, BigData 2024*. IEEE-Institute of Electrical and Electronics Engineers Inc., 2024.

[25] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, and Y. Zhou, "A hybrid approach to privacy-preserving federated learning," in *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, 2019, pp. 1–11.

[26] Y. Liu, Y. Liu, Z. Liu, Y. Liang, C. Meng, J. Zhang, and Y. Zheng, "Federated forest," *IEEE Transactions on Big Data*, 2020.

[27] A. Aminifar, F. Rabbi, K. I. Pun, and Y. Lamo, "Privacy preserving distributed extremely randomized trees," in *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, 2021, pp. 1102–1105.

[28] A. Aminifar, F. Rabbi, and Y. Lamo, "Scalable privacy-preserving distributed extremely randomized trees for structured data with multiple colluding parties," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2655–2659.

[29] A. Aminifar, M. Shokri, F. Rabbi, V. K. I. Pun, and Y. Lamo, "Extremely randomized trees with privacy preservation for distributed structured health data," *IEEE Access*, vol. 10, pp. 6010–6027, 2022.

[30] R. S. Antunes, C. André da Costa, A. Küderle, I. A. Yari, and B. Eskofier, "Federated learning for healthcare: Systematic review and architecture proposal," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 4, pp. 1–23, 2022.

[31] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of healthcare informatics research*, vol. 5, pp. 1–19, 2021.

[32] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[33] D. C. Nguyen, Q.-V. Pham, P. N. Pathirana, M. Ding, A. Seneviratne, Z. Lin, O. Dobre, and W.-J. Hwang, "Federated learning for smart healthcare: A survey," *ACM Computing Surveys (Csur)*, vol. 55, no. 3, pp. 1–37, 2022.

[34] W. H. Organization *et al.*, *Epilepsy: a public health imperative*. World Health Organization, 2019.

[35] World Health Organization, "Epilepsy Fact Sheet," https://www.who.int/news-room/fact-sheets/detail/epilepsy, 2024, [Accessed: Dec. 30, 2024].

[36] ——, "Epilepsy: Health Topic," https://www.who.int/health-topics/epilepsy#tab=tab_1, 2024, [Accessed: Dec. 30, 2024].

[37] D. J. Thurman, D. C. Hesdorffer, and J. A. French, "Sudden unexpected death in epilepsy: assessing the public health burden," *Epilepsia*, vol. 55, no. 10, pp. 1479–1485, 2014.

[38] E. Trinka, L. J. Rainer, C. A. Granbichler, G. Zimmermann, and M. Leitinger, "Mortality, and life expectancy in epilepsy and status epilepticus—current trends and future aspects," *Frontiers in Epidemiology*, vol. 3, p. 1081757, 2023.

[39] B. Huang, A. Abtahi, and A. Aminifar, "Lightweight machine learning for seizure detection on wearable devices," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–2.

[40] R. Zanetti, A. Aminifar, and D. Atienza, "Robust epileptic seizure detection on wearable systems with reduced false-alarm rate," in *2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC)*. IEEE, 2020, pp. 4248–4251.

[41] B. Huang, R. Zanetti, A. Abtahi, D. Atienza, and A. Aminifar, "Epilepsynet: Interpretable self-supervised seizure detection for low-power wearable systems," in *2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, 2023, pp. 1–5.

[42] A. Shahbazinia, F. Ponzina, J. A. Miranda, J. Dan, G. Ansaloni, and D. Atienza, "Resource-efficient continual learning for personalized online seizure detection," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2024, pp. 1–7.

[43] F. Forooghifar, A. Aminifar, and D. Atienza, "Resource-aware distributed epilepsy monitoring using self-awareness from edge to cloud," *IEEE transactions on biomedical circuits and systems*, 2019.

[44] F. Forooghifar, A. Aminifar, T. Teijeiro, A. Aminifar, J. Jeppesen, S. Beniczky, and D. Atienza, "Self-aware anomaly-detection for epilepsy monitoring on low-power wearable electrocardiographic devices," in *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, 2021, pp. 1–4.

[45] F. Forooghifar, A. Aminifar, L. Cammoun, I. Wisniewski, C. Ciumas, P. Ryvlin, and D. Atienza, "A self-aware epilepsy monitoring system for real-time epileptic seizure detection," *Mobile Networks and Applications*, pp. 1–14, 2022.

[46] F. Forooghifar, A. Aminifar, and D. A. Alonso, "Self-aware wearable systems in epileptic seizure detection," in *2018 21st Euromicro conference on digital system design (DSD)*. IEEE, 2018, pp. 426–432.

[47] S. Frey, M. A. Lucchini, V. Kartsch, T. M. Ingolfsson, A. H. Bernardi, M. Segessenmann, J. Osieleniec, S. Benatti, L. Benini, and A. Cossettini, "Gapses: Versatile smart glasses for comfortable and fully-dry acquisition and parallel ultra-low-power processing of eeg and eog," *arXiv preprint arXiv:2406.07903*, 2024.

[48] J. Dan, U. Pale, A. Amirshahi, W. Cappelletti, T. M. Ingolfsson, X. Wang, A. Cossettini, A. Bernini, L. Benini, S. Beniczky *et al.*, "Szcore: Seizure community open-source research evaluation framework for the validation of electroencephalography-based automated seizure detection algorithms," *Epilepsia*, 2024.

[49] D. Pascual, A. Aminifar, and D. Atienza, "A self-learning methodology for epileptic seizure detection with minimally-supervised edge labeling," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019, pp. 764–769.

[50] A. Amirshahi, A. Thomas, A. Aminifar, T. Rosing, and D. Atienza, "M2d2: Maximum-mean-discrepancy decoder for temporal localization of epileptic brain activities," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 202–214, 2022.

[51] A. Van de Vel, K. Cuppens, B. Bonroy, M. Milosevic, K. Jansen, S. Van Huffel, B. Vanrumste, P. Cras, L. Lagae, and B. Ceulemans, "Non-eeg seizure detection systems and potential sudep prevention: state of the art: review and update," *Seizure*, vol. 41, pp. 141–153, 2016.

[52] M. J. Morrell, "Responsive cortical stimulation for the treatment of medically intractable partial epilepsy," *Neurology*, vol. 77, no. 13, pp. 1295–1304, 2011.

[53] G. Rubboli, M. H. Bø, K. Alfstad, S. Armand Larsen, M. D. H. Jacobsen, M. Vlachou, S. Weisdorf, R. Rasmussen, A. Egge, O. Henning *et al.*, "Clinical utility of ultra long-term subcutaneous electroencephalographic monitoring in drug-resistant epilepsies: a "real world" pilot study," *Epilepsia*, vol. 65, no. 11, pp. 3265–3278, 2024.

[54] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. Siegelbaum, A. J. Hudspeth, S. Mack *et al.*, *Principles of neural science*. McGraw-hill New York, 2000, vol. 4.

[55] P. Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[56] R. B. Blackman and J. W. Tukey, "The measurement of power spectra from the point of view of communications engineering—part i," *Bell System Technical Journal*, vol. 37, no. 1, pp. 185–282, 1958.

[57] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza, "Real-time classification technique for early detection and prevention of myocardial infarction on wearable devices," in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2017.

[58] K. M. Tsiouris, V. C. Pezoulas, M. Zervakis, S. Konitsiotis, D. D. Koutsouris, and D. I. Fotiadis, "A long short-term memory deep learning network for the prediction of epileptic seizures using eeg signals," *Computers in biology and medicine*, vol. 99, pp. 24–37, 2018.

[59] K. M. Tsiouris, S. Konitsiotis, S. Markoula, G. Rigas, D. D. Koutsouris, and D. I. Fotiadis, "Unsupervised detection of epileptic seizures from eeg signals: A channel-specific analysis of long-term recordings," in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2018, pp. 92–95.

[60] A. Aminifar, B. Huang, A. Abtahi, and A. Aminifar, "Lightff: Lightweight inference for forward-forward algorithm," in *ECAI 2024*. IOS Press, 2024, pp. 1728–1735.

[61] C. Gómez, P. Arbeláez, M. Navarrete, C. Alvarado-Rojas, M. Le Van Quyen, and M. Valderrama, "Automatic seizure detection based on imaged-eeg signals through fully convolutional networks," *Scientific reports*, vol. 10, no. 1, p. 21833, 2020.

[62] A. Van Esbroeck, L. Smith, Z. Syed, S. Singh, and Z. Karam, "Multi-task seizure detection: addressing intra-patient variation in seizure morphologies," *Machine Learning*, vol. 102, pp. 309–321, 2016.

[63] M. Tacke, K. Janson, K. Vill, F. Heinen, L. Gerstl, K. Reiter, and I. Borggraefe, "Effects of a reduction of the number of electrodes in the eeg montage on the number of identified seizure patterns," *Scientific Reports*, vol. 12, no. 1, p. 4621, 2022.