Journal of Neural Engineering

PURPOSE-LED PUBLISHING™

ACCEPTED MANUSCRIPT

Select for better learning: Identifying high-quality training data for a multimodal cyclic transformer

To cite this article before publication: Jingwei Zhang *et al* 2025 *J. Neural Eng.* in press https://doi.org/10.1088/1741-2552/adbec0

Select for Better Learning: Identifying High-quality Training Data for A Multimodal Cyclic Transformer

Jingwei Zhang[1] †, **Zhaoyi Liu**[2] †, **Christos Chatzichristos**[1], **Sam Michiels**[2], **Wim Van Paesschen**[3,4], **Danny Hughes**[2], **Maarten De Vos**[1,5]

[1] STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, Department of Electrical Engineering, KU Leuven, Leuven, Belgium
[2]imec-Distrinet, Computer Science, KU Leuven, Leuven, Belgium
[3] Laboratory for Epilepsy Research, KU Leuven, Leuven, Belgium
[4] Reference center for refractory epilepsy, Department of Neurology, UZ Leuven, Leuven, Belgium
[5] Department of Development and Regeneration, KU Leuven, Leuven, Belgium

E-mail: jingwei.zhang@esat.kuleuven.be
Correspondence: Jingwei Zhang, Kasteelpark Arenberg 10, bus 2446, B-3001, Leuven, Belgium

Abstract. Tonic-clonic seizures (TCSs), which present a significant risk for sudden unexpected death in epilepsy (SUDEP), require accurate detection to enable effective long-term monitoring. Previous studies have demonstrated the advantages of multimodal seizure detection systems in reliably detecting TCSs over extended periods. However, the effectiveness of these data-driven systems depends heavily on the availability of reliable training data. To address this need, we propose an innovative data selection method designed to identify high-quality training samples. Our approach evaluates sample quality based on learning difficulty, classifying samples with lower learning difficulty as higher quality. We then introduce a confidence-based method to quantify the proportion of high-quality samples within the dataset. Using this data selection method, we develop a training pipeline that enhances the training process of multimodal seizure detection models. Experimental results show that our method improves the performance of a state-of-the-art TCS detection model by 11%.

1. Introduction

Tonic-clonic seizures (TCSs), which pose a significant risk for sudden unexpected death in epilepsy (SUDEP), are the most severe seizure type and can be of either focal or generalized onset [1]. Long-term seizure monitoring is crucial in improving the outcomes and quality of life for patients suffering from TCSs, as uncontrolled TCSs can lead to various complications, including physical injuries, cognitive impairments, and psychosocial dysfunctions [2].

† Equal Contribution

By systematically tracking the frequency, duration, and characteristics of TCSs, seizure monitoring provides valuable insights into disease progression and treatment efficacy [3, 4]. Although video-electroencephalogram (video-EEG) is considered the gold standard for seizure monitoring, its high cost and discomfort make it impractical for long-term use. In contrast, wearable electroencephalogram (EEG) devices offer a more accessible and less intrusive alternative, making them more suitable for continuous long-term seizure monitoring [5, 6, 7].

Long-term seizure monitoring is empowered by automated seizure detection methods that can be integrated into wearable devices [8, 9, 10, 11]. While our previous work has shown that (wearable) EEG is the primary modality for automated TCS detection, integrating multiple modalities such as electromyography (EMG) and accelerometry (ACC) significantly enhances multimodal detection methods [12, 13, 9, 14]. Multimodal TCS detection systems capture the different semiological phases of TCSs and leverage the strengths of each modality, leading to improved performance.

As most multimodal methods are data-driven, the availability of high quality data directly influences the performance of multimodal seizure detection models. Ideally, all modalities should be available, aligned, and free of noise within every sample. However, in practical applications, data quality frequently fails to meet ideal standards due to multiple factors [15].

This discrepancy arises from several challenges, the first of which is the presence of noisy labels in seizure monitoring datasets, attributed to inconsistencies and unreliability in annotations [16].Medical experts may find it challenging to accurately annotate seizures in non-standard modalities used for long-term monitoring due to insufficient training and knowledge. This lack of consistency and reliability may result in deviations from ground-truth annotations. Consequently, non-seizure signals and artifacts could be incorrectly labeled as seizures, introducing label noise into the dataset.

Secondly, the signal quality of the samples can be poor due to the artifacts and interference that obscure or mask seizure activity [8]. These corrupted samples introduce noise into the dataset, which in turn increases both bias and variance in models trained on them. This noise can mislead the model into prioritizing irrelevant features, thereby reducing the model's ability to generalize to unseen data [17].

Finally, the alignment of relevant information is not always consistent across different modalities. TCS patterns can shift across different modalities due to time delays between sensors. This misalignment makes it challenging to effectively leverage complementary information from different sources [18]. While well-aligned data enables models to better identify correlations and dependencies, such misalignment of important patterns can confuse models and decrease their performance.

Using data with suboptimal quality during training can diminish model performance by increasing bias and variance, reducing robustness, and leading to unreliable predictions, ultimately hindering the model's ability to generalize effectively to new data. When datasets are created we can assume there is ample of good quality data, and a fraction of lower quality data.

To reduce the effect of low-quality data, our objective is to develop a novel data selection method for our multimodal seizure detection method, the Multimodal Input Cyclic Transforme (MICT) proposed in [19], to ensure that ample high-quality data and a manageable fraction of lower-quality data are utilized during training.

It has been observed and demonstrated in numerous prior studies [20, 21] that deep neural networks exhibit a tendency to learn the easiest samples first and progressively memorize more challenging ones. As a result, under consistent training conditions, e.g., using the same network architecture, low-quality data produces higher training loss than high-quality data. This trend arises because noisy samples inherently deviate from the underlying data distribution, making them more challenging for the model to learn effectively.Inspired by this "memorization effect" in the learning process, we find a strong correlation between learning speed and data quality. Building on this insight, we introduce a concept of "*learning difficulty*" that quantifies the possibilities of a sample being "low quality" or hard to learn from based on its training loss. This metric helps in identifying and potentially excluding low-quality samples from the training dataset.

Given the "learning difficulty" of each training sample, we further automatically select data for training based on confident learning. This method is specifically designed for the MICT, aiming to select the optimal training data to train a robust and comprehensive seizure detection model.

In this paper, we introduce a model to measure the learning difficulty of each training sample. We then present a confidence-based method to quantify the proportion of high-quality training samples based on their associated learning confidence. Finally, leveraging the defined learning difficulty and high-quality sample quantification, we propose a learning pipeline that automatically selects the most reliable samples for training.

Experimental results showed the advantages of our training pipeline from various perspectives:

(i) **Advanced Data Quality Selection**: Our innovative method selects samples that are noise-free, complete, and uncorrupted, and enhances data quality by evaluating training loss to ensure only the highest quality data is used for training.

(ii) **Optimal Data Utilization**: To finely balance discarding low-quality data with retaining essential information, our method employed a confidence-based method that strategically assesses the prevalence of low-quality samples in our dataset.

The rest of this paper is organized as follows: In Section 2, we introduce the MICT, the current state-of-the-art (SOTA) model for TCS detection, as a preliminary foundation for this work. In Section 3, we propose a training pipeline designed to automatically select reliable training samples for MICT. Section 4 presents the evaluation metrics and experimental results assessing the performance of MICTs trained by different pipeline. Finally, we conclude the paper with a discussion in Section 5.

2. Related Work

Approaches to learning from low-quality or noisy data can be broadly divided into two categories: (1) Training sample selection: identifying high-quality samples from the noisy dataset for re-training; and (2) Designing robust model architectures or loss functions: allowing for the direct training of noise-robust models on the noisy dataset.

2.1. Training Sample Selection

To avoid negative effect of low-quality training samples, many recent studies [22, 23, 24, 25, 26, 27, 28] have adopted sample selection that selects high-quality samples form noisy training set.

Training sample selection is a well-founded and effective strategy; however, it is prone to cumulative errors stemming from incorrect selections, particularly when the training data contains many ambiguous classes. To address this issue, recent approaches often employ multiple deep neural networks (DNNs) that collaborate with each other [29]. Alternatively, studies, such as O2U-Net, have implemented cyclical training methods that adjust hyperparameters in a cyclical manner, transitioning between overfitting and underfitting. This approach facilitates the identification of noisy samples based on their normalized average training loss [23].

2.2. Robust Loss Function

Robust loss function is effective for reducing the negative impact of noisy labels by adjusting the loss of all training examples before updating the DNN [30, 31, 32]. To make the traditional optimization function (such as cross-entropy loss) more robust to low-quality data, the loss function is commonly designed in three strategies: (a) loss correction, which estimates the noise transition matrix to correct the forward or backward loss [30, 33]; (2) loss re-weighting, which assigns different importance to each sample for a weighted training approach [34]; (3) label refurbishment, which adjusts the loss using a refined label derived from a convex combination of noisy and predicted labels [35]. As a result, the update rule during training is adapted to minimize the adverse effects of label noise.

2.3. Robust Architecture

Various studies have introduced robust architectural modifications to model the noise transition matrix in noisy datasets [36, 37]. These changes usually included two types: (a) Design layer for noise adaptation, which is intended to mimic the label transition behavior in learning a DNN [38, 39]; (b) Robust regularization, which is used to enhance the generalization of DNNs by preventing overfitting [40, 41].

Notably, previous methods frequently rely on specific assumptions, such as noise distribution, the design of specialized loss functions, or the introduction of additional hyperparameters. These constraints can limit their adaptability and require

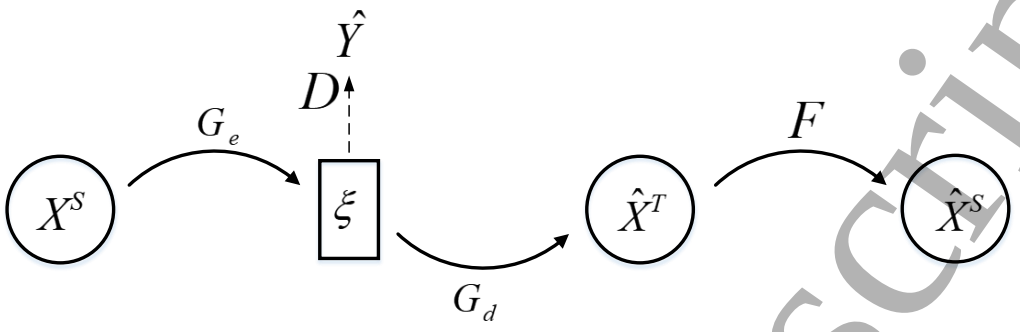


Figure 1: The general structure of the multimodal input cyclic transformer.

manual intervention for optimization. In contrast, our proposed approach is entirely hyperparameter-free, circumventing the need for such assumptions or custom loss functions. This inherent flexibility ensures robustness of proposed approach across a wide range of datasets and applications, eliminating the necessity for manual tuning.

3. Preliminaries: Cyclic Transformer

3.1. Model Architecture

In this section, we briefly review the MICT, a SOTA model for multimodal TC seizure detection detailed in our previous publication [19]. As shown in Figure 1, the MICT has an encoder-decoder-reconstructor architecture, designed to learn robust representations of TC seizures. Additionally, it also includes a classifier for accurate input predictions.

Given a positional encoded input sequence $X^S \in \mathbb{R}^{d_{seq} \times d_{feat}}$, the encoder $G_e(\cdot)$ first processes the input modalities through transformer blocks to embed them into intermediate representations, where d_{seq} and d_{feat} represent the length of input sequence and the dimension of the features respectively.

$$\xi = G_e\left(X^S\right) \tag{1}$$

Once the intermediate representation ξ is obtained, the transformer decoder $G_d(.)$ projects the intermediate representation ξ into the target modality sequence $\hat{X}^T$

$$\hat{X}^T = G_d\left(\xi\right) \tag{2}$$

To learn a reliable and noise-robust intermediate representation via the translation X^S to $\hat{X}^T$, we added a reconstructor $F(\cdot)$ to the architecture.

$$\hat{X}^S = F\left(\hat{X}^T\right) \tag{3}$$

This additional reconstructor enables back-translation [42, 43, 44, 45] which acts as an additional regularization for learning a robust intermediate representation. To

make sure the learned intermediate representations focus on the useful information for the final classification task, a classifier is integrated into the MICT's architecture.

The classifier $D(\cdot)$ takes the intermediate representation ξ as input and output the prediction $\hat{Y}$:

$$\hat{Y} = D\left(\xi\right) \tag{4}$$

3.2. Learning Objective

The cyclic transformer is trained on the paired multimodal data and labels.

The optimization objective includes two parts: reconstruction loss between two modalities and the final classification loss.

The first loss is the forward translation (mapping) loss L_f which is defined as the following expectation:

$$L_f = E\left(l_{X^T}(\hat{X}^T, X^T)\right) \tag{5}$$

where l_{X^T} denotes root mean square error, $E(\cdot)$ denotes the expectation calculation.

And the second loss is the cycle consistency loss which is defined as the following expectation:

$$L_b = E\left(l_{X^S}(\hat{X}^S, X^S)\right) \tag{6}$$

where l_{X^S} denotes root means square error here.

Then, the classification loss is defined as the following expectation:

$$L_c = E\left(l_c(\hat{Y}, Y)\right) \tag{7}$$

where l_c denotes cross entropy here.

The full objective is then:

$$L = L_f + L_b + L_c \tag{8}$$

4. Method

4.1. Dataset

The dataset was collected as part of the SeizeIT2 project [46]. The SeizeIT2 project adhered to the principles embodied in the Declaration of Helsinki and conformed to pertinent local statutory requirements. Written informed consent was obtained from all participants, or from their parent or legal guardian in the case of children under 16. The SeizeIT2 received approval from the ethical committee of UZ/KU Leuven, under the reference number S63631. Furthermore, the SeizeIT2 trial is registered under the identifier NCT04284072 on clinicaltrials.gov.

During the project, participants' BTE-EEG, EMG, ECG, and ACC signals were recorded by two Sensor Dot (SD) devices [47]. The first SD, known as the EEG SD, was positioned on the upper back of each participant to record BTE-EEG and ACC signals. The second SD, the EMG SD, was placed on the left chest to record EMG and

ECG signals. The BTE-EEG, EMG, and ECG were recorded with a sampling rate at 250 Hz, while ACC was recorded with a sampling rate at 25 Hz.

Among the participants, 27 participants had their TCSs recorded by both SDs, while 16 participants had their TCSs only recorded by the EEG SD.

As every recording has been carefully reviewed by the neurologists, there will not be miss-annotated seizures in the recordings.

Two datasets were created according to signal availability:

- Dataset A: Comprises 3233 hours of recordings from participants with TCSs recorded by both SDs. Dataset A recorded 43 focal to bilateral TCSs (FBTCS) and 1 generalized TCS (GTCS), with an average TCS duration of 124 seconds (range: 55–661 s)
- Dataset B: Comprises 1284 hours of recordings from participants with TCSs recorded only by the EEG SD. Dataset B recorded 19 FBTCS, with an average TCS duration of 183 seconds (range: 44–570 s).

4.2. Model Training Pipeline Based on Data Selection

Let's consider the training process of the multimodal MICT network. Suppose $S = \{Z_i = (X_i^S, X_i^T, Y_i), 1 \leq i \leq N\}$ is the data set with N training samples, where Z_i denotes the i-th training sample. The learning objective of a MICT is to optimize its parameters θ by minimizing the loss function $L(\theta, S)$.

Assuming the loss function $L(\theta, S)$ is twice-differentiable, MICT updates its parameters at each iteration t with a learning rate η,

$$\theta_{t+1} = \theta_t - \eta \nabla \left(\frac{1}{|S|} \sum_{\{Z_i\} \in S} L(\theta, S) \right) \tag{9}$$

As the quality of the data used for training has a significant effect on the learning process, our objective is to build a robust training pipeline for the optimization of the model's parameters. The strategy is to select the high-quality samples from the noisy training set to create a new training set S_h at each epoch. This ensures that the deep neural network $f(S; \theta)$ will only update its parameters θ with the high-quality samples,

$$\theta_{t+1} = \theta_t - \eta \nabla \left(\frac{1}{|S_h|} \sum_{\{Z_i\} \in S_h} L(\theta, S_h) \right) \tag{10}$$

4.3. Learning Difficulty Modeling

As shown in Algorithm 1, the model training pipeline selects high-quality samples based on their learning difficulties at each epoch. In this section, we will introduce our proposed metrics that can be used to estimate the learning difficulty of each sample.

Algorithm 1 Model Training Pipeline Based on Data Selection

Input: dataset S, network $f(S;\theta)$, training epoch T
Initialize selected training set $S_h = \emptyset$
Estimate the amount of high-quality samples $|S_{s_i}|$ in each class based on confident joint
Initialize a network $f(\theta)$
for $1 \leq i \leq T$ **do**
    if $i = 1$ **then**
        Train $f(S;\theta)$ on S
    else
        Train $f(S;\theta)$ on S_h
    end if
    for $z_i \in S$ **do**
        $Diffic_i \leftarrow Difficulty(f(z_i;\theta))$
    end for
    Rank samples with their learning difficulty $Diffic_i$
    Select first $|S_{s_i}|$ samples as high-quality set S_h in each class
end for

Deep neural networks initially prioritize learning simpler and more informative samples while gradually attempting to memorize more complex or noisy samples as training continues [20]. This results in lower training loss for the easier samples [48]. Low-quality samples, which are more challenging for the model to learn, usually exhibit higher training loss compared to high-quality samples [20, 21].

Given the observation of lower loss associated with easier samples, we define the learning difficulty of each sample by analyzing the different components of the training loss.

The classification loss indicates the difficulty in the classification task. Having a high classification loss suggests the sample is out of the distribution for its labeled class. Therefore, a sample with high classification loss is potentially an outlier. The forward translation loss indicates the difficulty in the forward translation task. Having a high translation loss suggests that the multiple modalities are not well aligned or one or more modalities have bad signal quality. The cycle consistency loss acts as a regularization technique and therefore does not directly inflect the learning difficulty of a sample.

Based on the aforementioned analysis, we define the learning difficulty by its classification loss and its forward translation loss:

$$Diffic_i = E\left(l_c(\hat{Y}, Y)\right) + E\left(l_{X^T}(\hat{X}^T, X^T)\right) \tag{11}$$

4.4. Confidence-Based Quantification of High-Quality samples

Given the learning difficulty of each sample, we can rank the training samples from very likely to be high-quality to very likely to be low-quality. Consequently, we can potentially improve the model's performance by only including the high-quality samples for model training. However, there is a trade-off between including too many samples that may still contain low-quality data and including too few samples, which could lead to the loss of valuable information. Therefore, an effective method is required to quantify the optimal number of samples to include for training.

Samples from minority classes usually have a higher learning difficulty compared to those from majority classes. Thus, selecting the samples with the smallest learning difficulty will lead to a biased selection towards the majority class.

To avoid a biased sample selection, we introduce a class-dependent sample selection which selects the samples individually within each class. Let S_{class_i} denote the samples of i-th class each class, we will select $|S_{s_i}|$ samples that with smaller learning difficulty from S_{class_i}, where $|S_{s_i}|$ denotes the number of selected samples in i-th class.

Considering that the noise distribution could be asymmetric across classes, a $|S_{s_i}|$ is supposed to be estimated for each class. The number of selected samples can be estimated as the amount of training samples that could be confidently identified as high quality in each class.

The confident joint matrix C can be used to estimated the $|S_{s_i}|$, where C is an $m \times m$ matrix and m represents the number of classes. Each element $C_{i,j}$ in C denotes the number of samples in Z that were labeled as class i but actually should belong to another class j. Therefore, the diagonal elements of C represents the number of clean samples in each class.

The value of $C_{i,j}$ can be estimated using the following equation:

$$C_{i,j} =| \hat{Z}_{\tilde{Y}=i,Y^*=j} | \tag{12}$$

where $\hat{Z}_{\tilde{Y}=i,Y^*=j}$ is the set of samples that were labeled as class i but should actually belong to class j.

The $\hat{Z}_{\tilde{Y}=i,Y^*=j}$ is estimated by running an internal cross-validation on the training set S. Traditionally, a sample is identified as belonging to a certain class when its predicted probability of belonging to that class exceeds a predefined per-class threshold [49]. However, in this study, which focuses on a binary classification scenario, each sample is classified exclusively as either one class or the other. In this context, no explicit threshold is required. Instead, a sample is assigned to a class if its predicted probability of belonging to that class is greater than its predicted probability of belonging to the alternative class.

Given the confident joint C, the number of selected samples $|S_{s_i}|$ can be estimated as corresponding diagonal element of the confident joint C.

With the confident joint C, the confident matrix $Q_{i,j}$, which offer a deeper insight of the distribution of data quality, can be estimated as,

$$Q_{i,j} = \frac{C_{i,j}}{\sum_{j\in[m]} C_{i,j}} |Z_{\tilde{Y}=i}| \times \left(\sum\nolimits_{i\in[m],j\in[m]} \left(\frac{C_{i,j}}{\sum_{j\in[m]} C_{i,j}} |Z_{\tilde{Y}=i}| \right) \right)^{-1} \quad (13)$$

where m denotes the number of classes in the dataset, $Z_{\tilde{y}=i}$ denotes the number of samples labeled as class i. The denominator normalizes the confidence matrix $Q_{i,j}$, ensuring that the sum of all its elements satisfies $\sum_{i\in[m]} \sum_{j\in[m]} Q_{i,j} = 1$ to guarantee that the matrix represents a valid probability distribution, where the total sum of all entries equals one.

5. Experiments and Results

5.1. Model Implementation

Two datasets were employed to evaluate the performance of the proposed approach. Based on the availability of modalities, Dataset A was used for both training and testing, while Dataset B was used solely for testing.

Firstly, a leave-one-participant-out cross-validation approach was conducted on Dataset A. In each fold, data from one participant was set aside as the test set to evaluate the model's performance, while data from the remaining participants was divided into a training set and a validation set. Importantly, each participant's data was used exclusively in either the training or validation set. Additionally, Dataset B served as an independent test set to further evaluate the models trained on Dataset A.

The input signals were cut into 60-second segments and were then transformed into time-frequency spectrogram $X \in \mathbb{R}^{d\times f}$, where d denotes the time indices and f denotes the frequency bins. The frequency range is from 3 Hz to 64 Hz for EEG and EMG, and 0 Hz to 12 Hz for ACC.

We used PyTorch to implement and train all the models [50]. The initial learning rate was set to 5×10^{-5} and then decayed by 0.75 at each epoch. The l2 regulation rate was set to 1×10^{-4}. The training process spanned 15 epochs, with the data presented in batches of 32. Early stopping [51], which selects the model with minimum validation loss, was applied in the training process.

5.2. Evaluation and Metrics

Term-based scoring metrics, which consider the overlap between ground-truth annotations and algorithm-generated hypothesis annotations, were employed for the evaluation. The true positives (TP), false negative (FN), and false positive (FP) were then defined as

- TP represents the detected TCS. It signifies that a seizure was detected when there were positive predictions within the duration of ground-truth annotation.
- FN is an undetected TCS.

- FP represents a continuous sequence of positive predictions that has no overlap with the ground-truth annotation. In other words, it refers to samples where the algorithm incorrectly identifies a seizure when there is no corresponding TCS in the ground-truth annotation

Typically, we expect the seizure detection system to detect as many seizures as possible while minimizing false positives. Therefore, sensitivity and FPR are used to assess performance.

Sensitivity is defined as the ratio of TPs to the sum of TPs and FNs:

$$Sensitivity = \frac{TPs}{TPs + FNs} \tag{14}$$

FPR is defined as the number of FPs within a unit of time:

$$FPR = \frac{FPs}{Duration} \tag{15}$$

Additionally, precision and the F1-score are also used to evaluate performance.

Precision is defined as the ratio of TPs to the sum of TPs and FPs:

$$Precision = \frac{TPs}{TPs + FPs} \tag{16}$$

F1-score is defined as a harmonic mean of sensitivity and precision:

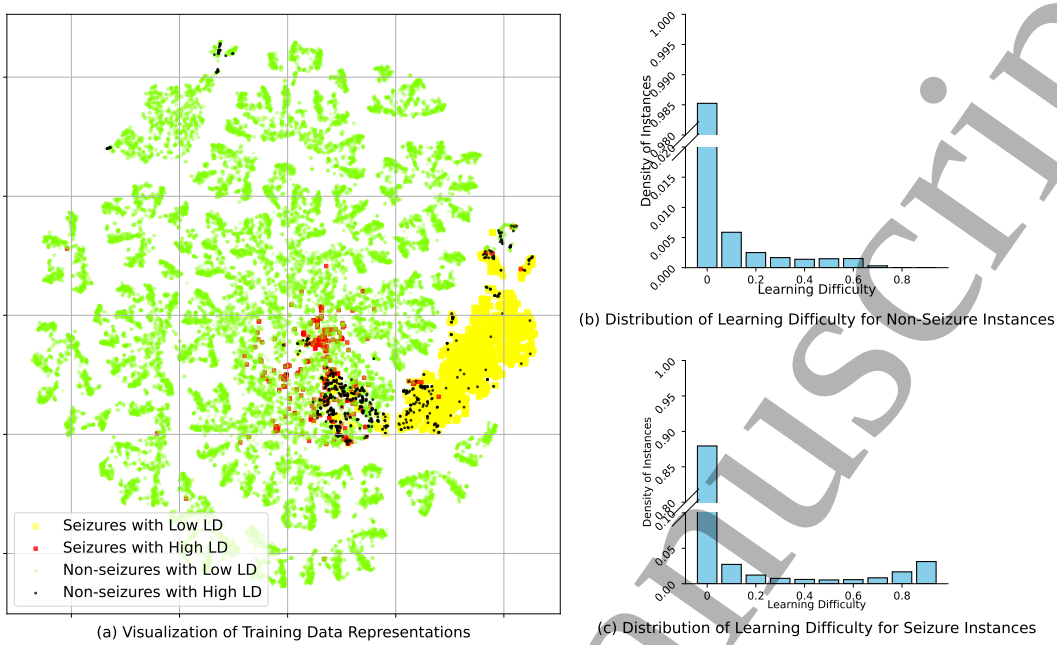
$$F1\text{-score} = \frac{Precision + Sensitivity}{2 \times Precision \times Sensitivity} \tag{17}$$

To ensure a fair evaluation of model performance, we opted not to compute sensitivity by averaging participant-specific sensitivities, as this approach could introduce biases due to the unequal distribution of seizures across patients. Instead, we sum the TPs, FPs , and FNs across all folds to calculate the overall sensitivity. This dataset-level approach provides a more accurate reflection of model performance across the entire population without being affected by seizure distribution disparities among patients.

5.3. Insights from Learning Difficulty

We illustrate the intermediate representations of the training data given by the MICT by embedding them into a two-dimensional space using t-SNE [52]. As shown in Figure 2, seizure samples and non-seizure samples are marked in different colors. The samples with the highest learning difficulty in the seizure class and non-seizure class are marked in red and black, respectively.

Samples with high learning difficulty are found mostly outside the main distribution. Many of them are located at the borders of the two classes. These samples are challenging to classify, complicating the training of the seizure detection model and potentially diminishing its generalization capabilities. Some of these samples are even



(a) Visualization of Training Data Representations

(b) Distribution of Learning Difficulty for Non-Seizure Instances

(c) Distribution of Learning Difficulty for Seizure Instances

Figure 2: Visualization of the intermediate representation distribution resulting from the MICT. (a) The distribution of two-dimensional tSNE of the representations of the input data given by the MICT. The green points and yellow points represent samples with low learning difficulty (LD) that are selected for training. The red points and black points represent samples with high learning difficulty that are not selected for training. (b) Distribution of learning difficulty for non-seizure samples. (c) Distribution of learning difficulty for seizure samples.

located within the main distribution of the opposite class, which may confuse the model during training and subsequently reduce its performance.

We further visualized the distribution of the learning difficulty for training data. As shown in Figure 2 (b) and Figure 2 (c), most of the training samples, particularly for the samples from the non-seizure class, have very low learning difficulty. Therefore, most samples will remain and contribute to the model's training when we remove the samples with high learning difficulty.

Table 1: Performance evaluation of the multimodal input cyclic transformer (MICT) trained by various data selection methods

	Metrics							
	Dataset A				Dataset B			
Model	Sens	FPR/24h	PPV	F1-score	Sens	FPR/24h	PPV	F1-score
Original Training Set	100.0%	0.4	43.6%	60.9%	84.2%	0.6	34.8%	49.2%
LD + GMM	100.0%	0.4	46.9%	63.8%	84.2%	0.9	24.2%	37.6%
CL	100.0%	0.5	41.5%	58.7%	78.9%	0.4	38.5%	51.7%
LD + CL	**100.0%**	**0.3**	**51.2%**	**67.7%**	**84.2%**	**0.3**	**48.5%**	**61.5%**

Abbreviations: Sens, sensitivity; FPR/24h, false positive rate per 24 hours; PPV, positive predictive value (precision); LD, learning difficulty;GMM, Gaussian Mixture Model; CL, confident learning.

5.4. Confidence Learning for the Quantification of High-Quality Samples

We quantified the number of high quality samples by estimating the confident joint of the dataset. The confident joint is estimated by confident learning, which counts the number of training samples that have high confidence in their predictions [49]. This approach helps in determining the optimal number of high-quality training samples to retain, thereby improving the model's performance and robustness.

Furthermore, since the number of training samples varies across each fold in cross-validation, we also calculate the selection ratio to gain a better insight into the consistency of the proportion of selected samples. Since each row of $Q_{i,j}$ represents the data distribution for a specific class, the selection ratio for a certain class is obtained by dividing the diagonal element of that row by the sum of all elements in the same row. For example, the confidence matrix for the first fold during cross-validation on Dataset A is $\begin{bmatrix} 0.084 & 0.008 \\ 0.012 & 0.896 \end{bmatrix}$. Based on the first row of this matrix, the selection ratio for the seizure class is calculated to be 91.3%. Since the results for Dataset A were obtained using leave-one-patient-out cross-validation, the confidence matrices vary across folds. To address this variability, we report the median and variance of the selection ratios. Specifically, the median selection ratio is 90.8%, with a variance of 9.8×10^{-5}.These statistics demonstrate that the selection ratio remains consistent across folds, underscoring the robustness of the approach.

Following the estimation confident joint, training samples can be selected by ranking them based on their learning difficulty. This modular approach allows us to select high-quality training samples while ensuring that low-quality samples are removed. This approach ensures that only the most reliable samples are used for training, which can enhance the model's overall performance and robustness.

To compare the effectiveness of different methods for quantifying high-quality samples, we also applied a Gaussian Mixture Model (GMM) to estimate the number of high learning difficulty samples in the dataset. The GMM first fits the learning difficulty distribution of the training samples and then separates them into two clusters.

The cluster with lower learning difficulty is identified as high-quality samples and used for training.

To evaluate the effectiveness of using high-quality samples for model training, we conducted experiments to compare the performance of the MICT trained on the selected training dataset to that trained on the original dataset.

Table 1 presents the performance of the MICT model trained on various datasets. Experimental results show that the model trained on the original dataset achieved a sensitivity of 100%, an FPR of 0.4 per 24 hours, and an F1-score of 60.9% on Dataset A, and a sensitivity of 84.2%, an FPR of 0.6 per 24 hours, and an F1-score of 49.2% on Dataset B. In contrast, the MICT model trained on data selected using a combination of learning difficulty and confidence learning outperformed all other models. Specifically, it achieved a sensitivity of 100%, an FPR of 0.3 per 24 hours, and an F1-score of 67.7% on Dataset A, as well as a sensitivity of 84.2%, an FPR of 0.3 per 24 hours, and an F1-score of 61.5% on Dataset B.

While the aforementioned metrics are calculated from cumulative values of TPs, FNs, and FPs across all patients in each datasets, we further computed participant-specific metrics for each participant to perform a more granular statistical analysis. This allowed us to calculate p-values for both FPR and F1-score. Specifically, the p-values are 0.05 for FPR on Dataset A and 0.12 for FPR on Dataset B, while for the F1-score, the values are 0.16 for Dataset A and 0.10 for Dataset B. Although the results for Dataset B are not statistically significant, this is primarily due to the fact that, for many individual participants, the model already achieved an FPR of 0 and an F1-score of 1. Consequently, the final statistical t-test did not reach significance due to the relatively small number of subjects who exhibited improvement. However, the consistency of improvement across subjects and the large benefit for certain subjects reinforces the validity of our method.

Better performance is achieved when the confidence learning is used for quantification compared to GMM. This suggests that the confidence learning maintains a more effective balance by minimizing confusing samples while preserving the valuable information needed for robust model training.

5.5. Ablation Study: Confidence Learning as An Independent Data Selection Method

The confidence learning can also be directly used for training samples selection [49]. This method involves selecting samples where the model has high confidence in its predictions. Given the training set S, a sample is selected if its predicted label matches its observed label.

The performance of the seizure detection model trained on the samples selected by confident learning is listed in Table 1. Clearly, directly selecting training samples using confident learning does not improve the model's performance. This suggests that learning difficulty is a better metric for measuring the possibility of a training sample being noisy.

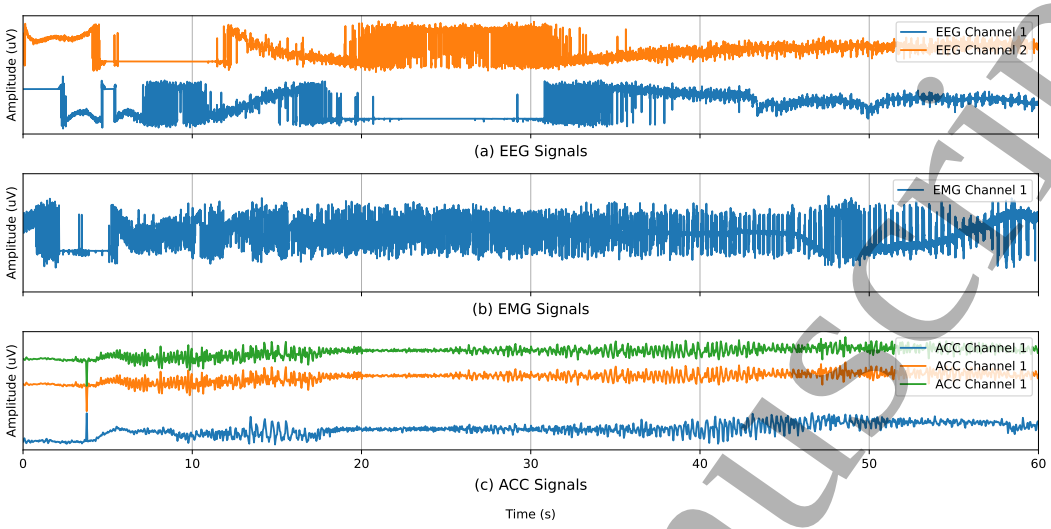


Figure 3: Visualization of Signals from a training sample characterized by high learning difficulty. (a) The EEG signals from the high learning difficulty sample with the first channel experiencing corruption between 5 seconds and 12 seconds, and the second channel experiencing corruption between 20 seconds and 31 seconds. (b) The EMG signals from the high learning difficulty sample with corruption between 2 second and 5 second. (c) The ACC signals from the high learning difficulty sample.

6. Discussion and Conclusion

In this paper, we proposed a data selection method designed for our multimodal seizure detection model called Multimodal Input Cyclic Transformer (MICT). Our method has shown its ability to enhance the quality of biomedical data, which often faces challenges related to reliability due to high variability, low completeness, and unavoidable noise and artifacts. The proposed method improves the training data quality through two key steps: 1) Utilizing learning difficulty for data-quality characterization, effectively ranking the training data from low quality to high quality, and 2) Employing a confidence-based method for data selection to estimate how many training samples should be included in the selected training dataset.

Through these steps, our method not only addresses the challenges of characterizing data quality but also optimizes the trade-off between removing samples that confuse the model and retaining useful information that enhances training effectiveness. We validated the effectiveness of this approach on two real-world seizure monitoring datasets, applying the proposed data selection method to the SOTA MICT model. This led to significant improvements in seizure detection performance on both datasets.

The proposed method automatically enhances the quality of training data, which is critical for the robustness of wearable seizure detection models. Such models often encounter unreliable data, so improving data quality is vital for their trustworthiness. Manual review of seizure monitoring data demands clinical expertise and is time-

consuming; therefore, automated data selection methods that improve data reliability are indispensable for effective seizure detection.

As shown in Figure 2, a visualization of learning difficulty distribution reveals that samples with high learning difficulty often lie outside the main data distribution or within the distribution of the opposite class, suggesting that low-quality samples include both outliers and potentially mis-annotated data. This supports the value of the learning difficulty metric in identifying label noise and data bias. Figure 3 illustrates a seizure sample with high learning difficulty, where EEG and EMG signals show significant corruption. Such visualizations demonstrate our method's effectiveness in detecting low-quality data, preventing the MICT model from being influenced by poor-quality training samples and improving overall model accuracy and robustness.

Additionally, our data selection method operates independently of hyperparameters, such as the threshold commonly required to separate low- and high-quality data. Unlike other approaches that necessitate dataset-specific hyperparameters adjustments [53, 54], our method uses a confidence-based method to automatically determine the quantity of high-quality data. This adaptability ensures that our approach remains robust across various datasets and applications, eliminating the need for manual tuning and streamlining the data selection process for better efficiency and accuracy.

In conclusion, we introduced a new data selection method for identifying reliable training data for the MICT model. We presented the learning difficulty metric to quantify training data quality and employed a confidence-based method to estimate the optimal size of the training set. Experimental results demonstrate that our method yields substantial improvements in the performance of SOTA TCS seizure detection models, underscoring its potential to advance the reliability and effectiveness of biomedical data applications.

7. Acknowledgment

The authors would like to thank all the patients who participated in this research. This study was funded by the EIT Health Grant: 21263 SeizeIT2 (Discreet Personalized Epileptic Seizure Detection Device). The research also received support from the following projects: FWO SB Project "Supporting the Development of Self-Regulation in Infants: A Promising Strategy in Preventive Mental Health Care" (S003524N), FWO Research Project "Artificial Intelligence (AI) for Data-Driven Personalized Medicine" (G0C9623N), FWO Research Project "Deep, Personalized Epileptic Seizure Detection" (G0D8321N), and the Bijzonder Onderzoeksfonds KU Leuven (BOF) project "Prevalentie van epilepsie en slaapstoornissen in de ziekte van Alzheimer" (C24/18/097). Additionally, this research was funded by the Flemish Government's AI Research Program.

Reference

[1] Philippe Ryvlin, Lina Nashef, Samden D Lhatoo, Lisa M Bateman, Jonathan Bird, Andrew Bleasel, Paul Boon, Arielle Crespel, Barbara A Dworetzky, Hans Høgenhaven, et al. Incidence and mechanisms of cardiorespiratory arrests in epilepsy monitoring units (mortemus): a retrospective study. *The Lancet Neurology*, 12(10):966–977, 2013.

[2] Michael R Sperling. The consequences of uncontrolled epilepsy. *CNS spectrums*, 9(2):98–109, 2004.

[3] Gregory D Cascino. Clinical indications and diagnostic yield of video-electroencephalographic monitoring in patients with seizures and spells. In *Mayo Clinic Proceedings*, volume 77, pages 1111–1120. Elsevier, 2002.

[4] Selim R Benbadis, Edward O'Neill, William O Tatum, and Leanne Heriaud. Outcome of prolonged video-eeg monitoring at a typical referral epilepsy center. *Epilepsia*, 45(9):1150–1153, 2004.

[5] Stefan Debener, Reiner Emkes, Maarten De Vos, and Martin Bleichner. Unobtrusive ambulatory eeg using a smartphone and flexible printed electrodes around the ear. *Scientific reports*, 5(1):16743, 2015.

[6] Ying Gu, Evy Cleeren, Jonathan Dan, Kasper Claes, Wim Van Paesschen, Sabine Van Huffel, and Borbála Hunyadi. Comparison between scalp eeg and behind-the-ear eeg for development of a wearable seizure detection system for patients with focal epilepsy. *Sensors*, 18(1):29, 2017.

[7] Steven Boeckx, Wim Van Paesschen, Brecht Bonte, and Jonathan Dan. Live demonstration: Seizeit-a wearable multimodal epileptic seizure detection device. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–1. IEEE, 2018.

[8] Lauren Swinnen, Christos Chatzichristos, Katrien Jansen, Lieven Lagae, Chantal Depondt, Laura Seynaeve, Evelien Vancaester, Annelies Van Dycke, Jaiver Macea, Kaat Vandecasteele, et al. Accurate detection of typical absence seizures in adults and children using a two-channel electroencephalographic wearable behind the ears. *Epilepsia*, 62(11):2741–2752, 2021.

[9] Jianbin Tang, Rima El Atrache, Shuang Yu, Umar Asif, Michele Jackson, Subhrajit Roy, Mahtab Mirmomeni, Sarah Cantley, Theodore Sheehan, Sarah Schubach, et al. Seizure detection using wearable sensors and machine learning: Setting a benchmark. *Epilepsia*, 62(8):1807–1819, 2021.

[10] Vaidehi Naganur, Shobi Sivathamboo, Zhibin Chen, Shitanshu Kusmakar, Ana Antonic-Baker, Terence J O'Brien, and Patrick Kwan. Automated seizure detection with noninvasive wearable devices: a systematic review and meta-analysis. *Epilepsia*, 63(8):1930–1941, 2022.

[11] Miguel Bhagubai, Lauren Swinnen, Evy Cleeren, Wim Van Paesschen, Maarten De Vos, and Christos Chatzichristos. Towards automated seizure detection with wearable eeg – grand challenge. *IEEE Open Journal of Signal Processing*, 5:717–724, 2024.

[12] Milica Milošević, Anouk Van de Vel, Bert Bonroy, Berten Ceulemans, Lieven Lagae, Bart Vanrumste, and Sabine Van Huffel. Detection of epileptic convulsions from accelerometry signals through machine learning approach. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2014.

[13] Milica Milošević, Anouk Van de Vel, Bert Bonroy, Berten Ceulemans, Lieven Lagae, Bart Vanrumste, and Sabine Van Huffel. Automated detection of tonic–clonic seizures using 3-d accelerometry and surface electromyography in pediatric patients. *IEEE journal of biomedical and health informatics*, 20(5):1333–1341, 2015.

[14] Jingwei Zhang, Lauren Swinnen, Christos Chatzichristos, Victoria Broux, Renee Proost, Katrien Jansen, Benno Mahler, Nicolas Zabler, Nino Epitashvilli, Matthias Duempelmann, et al. Multimodal wearable eeg, emg and accelerometry measurements improve the accuracy of tonic-clonic seizure detection. *Physiological Measurement*, 2024.

[15] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239, 2022.

[16] Jingwei Zhang, Christos Chatzichristos, Kaat Vandecasteele, Lauren Swinnen, Victoria Broux,

Evy Cleeren, Wim Van Paesschen, and Maarten De Vos. Automatic annotation correction for wearable eeg based epileptic seizure detection. *Journal of Neural Engineering*, 19(1):016038, 2022.

[17] Nick Seeuws, Maarten De Vos, and Alexander Bertrand. Electrocardiogram quality assessment using unsupervised deep learning. *IEEE Transactions on Biomedical Engineering*, 69(2):882–893, 2021.

[18] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. Learning alignment for multimodal emotion recognition from speech. *arXiv preprint arXiv:1909.05645*, 2019.

[19] Jingwei Zhang, Lauren Swinnen, Christos Chatzichristos, Wim Van Paesschen, and Maarten De Vos. Learning robust representations of tonic-clonic seizures with cyclic transformer. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[20] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

[21] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

[22] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR, 2018.

[23] Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. O2u-net: A simple noisy label detection approach for deep neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3326–3334, 2019.

[24] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International conference on machine learning*, pages 7164–7173. PMLR, 2019.

[25] Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pages 5739–5748. PMLR, 2019.

[26] Yueming Lyu and Ivor W Tsang. Curriculum loss: Robust learning and generalization against label corruption. In *International Conference on Learning Representations*.

[27] Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical image analysis*, 65:101759, 2020.

[28] Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24070–24079, 2023.

[29] Bob D de Vos, Gino E Jansen, and Ivana Išgum. Stochastic co-teaching for training neural networks with unknown levels of label noise. *Scientific reports*, 13(1):16875, 2023.

[30] Paul Albert, Eric Arazo, Tarun Krishna, Noel E O'Connor, and Kevin McGuinness. Is your noise correction noisy? pls: Robustness to label noise with two stage detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 118–127, 2023.

[31] Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. Noisebox: Towards more efficient and effective learning with noisy labels. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[32] Brandon Smart and Gustavo Carneiro. Bootstrapping the relationship between images and their clean and noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5344–5354, 2023.

[33] Masaki Kashiwagi, Keisuke Maeda, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Enhancing noisy label learning via unsupervised contrastive loss with label correction based on prior knowledge. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and*

Signal Processing (ICASSP), pages 6235–6239. IEEE, 2024.

[34] Erik Englesson and Hossein Azizpour. Robust classification via regression for learning with noisy labels. In *ICLR 2024-The Twelfth International Conference on Learning Representations, Messe Wien Exhibition and Congress Center, Vienna, Austria, May 7-11t, 2024*, 2024.

[35] Baharan Mirzasoleiman, Kaidi Cao, and Jure Leskovec. Coresets for robust training of deep neural networks against noisy labels. *Advances in Neural Information Processing Systems*, 33:11465–11477, 2020.

[36] Cheng-En Wu, Yu Tian, Haichao Yu, Heng Wang, Pedro Morgado, Yu Hen Hu, and Linjie Yang. Why is prompt tuning for vision-language models robust to noisy labels? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15488–15497, 2023.

[37] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. *IEEE Transactions on Image Processing*, 30:5984–5996, 2021.

[38] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1431–1439, 2015.

[39] Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. In *International Conference on Machine Learning*, pages 14153–14172. PMLR, 2022.

[40] Xiaobo Xia, Bo Han, Nannan Wang, Jiankang Deng, Jiatong Li, Yinian Mao, and Tongliang Liu. Extended t t: Learning with mixed closed-set and open-set noisy labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3047–3058, 2022.

[41] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*, 2020.

[42] Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. Neural machine translation with reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[43] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*, 2018.

[44] Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. Style transfer as unsupervised machine translation. *arXiv preprint arXiv:1808.07894*, 2018.

[45] Daimeng Wei, Zhanglin Wu, Hengchao Shang, Zongyao Li, Minghan Wang, Jiaxin Guo, Xiaoyu Chen, Zhengzhe Yu, and Hao Yang. Text style transfer back-translation. *arXiv preprint arXiv:2306.01318*, 2023.

[46] EIT-Health. Seizeit2. [online]. `https://eithealth.eu/product-service/seizeit2/`.

[47] Byteflies. Personalized wearable seizure monitoring. [online]. `https://www.byteflies.com/our-solutions`.

[48] Yingsong Huang, Bing Bai, Shengwei Zhao, Kun Bai, and Fei Wang. Uncertainty-aware learning against label noise on imbalanced datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6960–6969, 2022.

[49] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

[50] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[51] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002.

[52] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[53] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi,

Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*, 2020.

[54] Elisabeth RM Heremans, Nabeel Seedat, Bertien Buyse, Dries Testelmans, Mihaela van der Schaar, and Maarten De Vos. U-pass: An uncertainty-guided deep learning pipeline for automated sleep staging. *Computers in Biology and Medicine*, 171:108205, 2024.