iEDeaL: A Deep Learning Framework for Detecting Highly Imbalanced Interictal Epileptiform Discharges

Qitong Wang
LIPADE, Université Paris Cité
qitong.wang@etu.u-paris.fr

Stephen Whitmarsh
Sorbonne Université, Paris Brain Institute - ICM, Inserm, CNRS, APHP, Pitié-Salpêtrière Hospital
stephen.whitmarsh@icm-institute.org

Vincent Navarro
Sorbonne Université, Paris Brain Institute - ICM, Inserm, CNRS, APHP, Pitié-Salpêtrière Hospital
vincent.navarro@aphp.fr

Themis Palpanas
LIPADE, Université Paris Cité &
French University Institute (IUF)
themis@mi.parisdescartes.fr

ABSTRACT

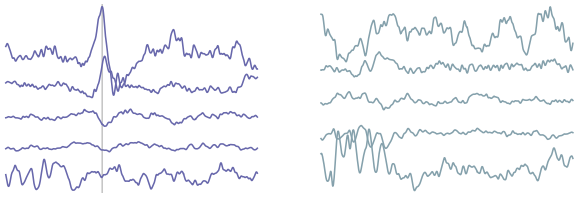
Epilepsy is a chronic neurological disease, ranked as the second most burdensome neurological disorder worldwide. Detecting Interictal Epileptiform Discharges (IEDs) is among the most important clinician operations to support epilepsy diagnosis, rendering automatic IED detection based on electroencephalography (EEG) signals an important topic. However, most existing solutions were designed and evaluated upon artificially balanced IED datasets, which do not conform to the real-world highly imbalanced scenarios. In this work, we propose the iEDeaL framework for automatic IED detection in challenging real-world use cases. The main components of iEDeaL are the new SC neural network architecture, to efficiently detect IEDs on raw EEG series instead of extracted features, and SaSu, a novel loss function to train SC by optimizing the F_β-score. Experiments on two real-world imbalanced IED datasets verify the advantages of iEDeaL in offering more accurate and efficient IED detection when compared with other state-of-the-art deep learning-based and spectrogram feature-based solutions.

PVLDB Reference Format:
Qitong Wang, Stephen Whitmarsh, Vincent Navarro, and Themis Palpanas. iEDeaL: A Deep Learning Framework for Detecting Highly Imbalanced Interictal Epileptiform Discharges. PVLDB, 16(3): 480 - 490, 2022.
doi:10.14778/3570690.3570698

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/qtwang/iEDeaL.

1 INTRODUCTION

[Motivation] Epilepsy is a neurological disorder characterized by disabling repetitive seizures, ranked as the second most burdensome neurological disorder worldwide [14]. Approximately 15% of epileptic patients are pharmacoresistant, for whom surgical treatment can be considered, requiring the identification of the brain area in which seizures are initiated. If no cortical lesion can be identified with neuroimaging (e.g. by magnetic resonance imaging, MRI), further EEG exploration *within* the brain is necessary by means of stereotaxic Electroencephalography (EEG) [105]. EEG provides crucial information on epileptic seizures and Interictal Epileptic Discharges (IEDs). IEDs are brief (<250 ms), morphologically defined events observed in the EEG of patients predisposed to spontaneous seizures of focal onset [81]. Figure 1 illustrates two subsequence examples of a multivariate EEG signal (derived using five sensors, or channels): one that contains an IED, and one that does not. Interictal spikes are highly correlated with spontaneous seizures [15, 58, 87]. The presence of IEDs is therefore used to support the diagnosis of epilepsy [62, 81, 94].



(a) IED series (b) Non-IED series

Figure 1: Examples of the IED and non-IED series in the ICM dataset. The gray line indicates the IED peak timestamp.

The detection and annotation of IEDs are extremely tedious and time-consuming for human experts. Furthermore, inter-expert and intra-expert agreement is known to be sub-optimal [2, 5, 30, 31, 36, 41, 80, 82, 86, 91, 92]. Efficient IED detection algorithms are of great potential utility for clinical and fundamental research of long-term EEG recordings by decreasing time-consuming manual annotation of IEDs, and increasing objectivity and sensitivity in their detection. Without automatic IED detection, the annotation of weeks of data typically recorded during clinical evaluations is practically impossible, limiting their usefulness and the impact of scientific studies, where long periodic patterns in epileptic activity spanning multi-day periods are observed [6, 74].

[Challenges] While early automatic IED detection algorithms might have been less accurate than that of experts, more recent implementations have been greatly improved [9, 61, 78, 80]. Nevertheless, automatic spike detection remains challenging for a number of reasons: definitions of a spike are simplistic, human experts often do not mark the same events as spikes, the ratio of candidate

Proceedings of the VLDB Endowment, Vol. 16, No. 3 ISSN 2150-8097.
doi:10.14778/3570690.3570698

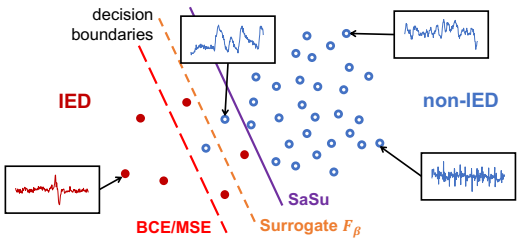


Figure 2: Influence on the classification decision boundaries of the common BCE/MSE loss functions, the surrogate F_β-score loss functions, and the proposed SaSu loss function.

spike events to actual spike events is very large (due to noise and artifacts), spike morphology and background vary widely between patients, and well-defined training sets are very time-consuming and expensive to develop [31, 93].

Traditional automatic IED detection algorithms are mostly designed upon expert-crafted feature templates [38, 45]. Recent years have witnessed the development of deep learning-based IED detection algorithms [76]. However, deep learning algorithms for IED detection encounter a specific methodological problem: IEDs are short-lasting and occur only sporadically. As a consequence, real-world datasets of IEDs contain only a small proportion of positive (IED) instances, compared to a much larger set of negative (non-IED) background. More often than not, the imbalanced nature of the data is not addressed when automatic IED detection algorithms are evaluated, and are instead tested on a comparable number of IED and non-IED instances. While this avoids the problem altogether [18], it prevents an evaluation of the generalizability of the models when applied to real-life imbalanced scenarios.

[Proposed Approach] In this paper, we propose iEDeaL (**i**mbalanced **IED** detection via **De**ep **L**earning), a deep learning approach for IED detection in real use cases. Our approach offers a practical, sensitive, and time-efficient procedure that requires only a small subset of the data to be annotated by experts. The model can then be used to identify IEDs in the whole dataset, thus, greatly increasing the return on human effort and expertise. The iEDeaL approach consists of a novel deep learning architecture, called Seanet Classifier (SC) to detect IED on raw EEG series, and SaSu (**Sa**mple-based F_β-score **Su**rrogate), a new loss function to guide the SC model training at optimizing the F_β-score. We note that compared to State-Of-The-Art (SOTA) solutions deployed in real-world IED detection [28, 76], raw EEG series-based detection removes the need for spectrogram generation, and hence achieves significantly faster detection.

The SC architecture is designed to efficiently detect IEDs based on the raw EEG series. It is derived from the SEAnet architecture, which has been verified to be effective for data series applications [89]. SC's basic structure follows an unmasked full-preactivation ResNet [33], augmented with exponentially increasing dilations [4]. Compared to recent SOTA IED detection methods based on spectrogram features [28, 76], SC is significantly faster, since it avoids spectrogram generation.

The SaSu loss function is a novel sample-based surrogate F_β-score loss function. (It is necessary to use a surrogate since the F_β-score is both *non-differentiable* and *non-decomposable*, preventing its direct deployment as a loss function in gradient-based learning algorithms for training neural network models.) Compared to other surrogate F_β-score loss functions [39, 46], SaSu further refines the decision boundary by drawing a portion of non-IED samples for training, while keeping track of the truth imbalance ratio. We illustrate this intuition in Figure 2. Compared to the Binary Cross Entropy (BCE) and Mean Squared Error (MSE) loss functions [29], surrogate F_β-score loss functions adjust the decision boundary by preventing its excessive expansion towards the minority class driven by the majority-dominated optimization. By considering a proper portion of non-IED samples, SaSu further refines the decision boundary to better fit the underlying distributions of both classes. Our analysis shows that when optimizing SaSu we converge at the same parameters as directly optimizing the F_β-score.

Employing SaSu on imbalanced datasets introduces the problem of imbalanced gradients from IED and non-IED instances, a problem common to surrogate F_β-score loss functions [39, 46]. As a result, the model is prone to gradient explosion/vanishing, hindering the SC model convergence. To deliver easy SC model training on SaSu, we equip iEDeaL with a dynamically weighted regularization, and two novel curriculum learning-based [7] auxiliary training strategies. The regularization reduces SaSu's gradients for well-classified instances [50]. The two auxiliary training strategies incorporate pretraining SC with BCE, and increasing the number of negative samples, starting from a balanced ratio and going up to the actual imbalanced ratio as the training epochs increase.

In summary, iEDeaL handles imbalanced datasets more accurately and 2 orders of magnitude faster than the SOTA methods in real-world IED detection. Our contributions are as follows:

(1) Our work is (to the best of our knowledge) the first to explicitly tackle the imbalance problem of real-world IED detection, by using surrogate F_β-score loss functions. Thus, the proposed approach enables analysts to train models that better fit their real datasets, leading to superior performance on the test data.

(2) We propose the iEDeaL framework to detect IEDs on imbalanced real-world datasets. iEDeaL is based on the new raw EEG series-based SC architecture and the novel SaSu loss function, serving as a general and easy-to-use framework.

(3) We design SaSu, a novel negative sampling-based surrogate F_β-score loss function. Its correctness is established upon approximately sharing the same condition of first-order stationary points with F_β-score.

(4) To facilitate SC model training on SaSu, we equip iEDeaL with simple auxiliary designs, i.e., a dynamically weighted regularization and two novel curriculum learning-based training strategies.

(5) Experimental results on two real-world IED datasets verified iEDeaL's effectiveness and efficiency by outperforming existing STOA deep learning and spectrogram feature-based solutions.

2 RELATED WORK

[IED detection] Traditional IED detection solutions are mostly expert-tuned template matching [43], or handcrafted feature engineering [32], including adaptive morphological filters [45] and signal envelope distribution modelling [38]. Motivated by its success in computer vision and natural language processing, employing deep learning techniques for IED detection has become popular and demonstrates promising performance [12]. Recent proposals cover

Convolutional Neural Networks (CNN) [1, 16, 25, 26, 54, 76, 85], Long Short-Term Memory (LSTM) [57] and Generative Adversarial Networks (GAN) [27, 84]. According to recent studies on data series classification, CNN models (e.g., ResNet [22] and InceptionTime [23]) generally achieve the SOTA performance [22]. Following the conventions, we also formalize IED detection to be a binary classification problem over sliding windows in our work.

However, existing deep models were trained with BCE or MSE loss functions, despite the fact that IED datasets are intrinsically imbalanced, leading to suboptimal performance in their real-world deployments. Moreover, their scalability was also limited by the time-consuming spectrogram feature extraction step [28, 76].

[Imbalanced Data Series Classification] In general machine learning studies, mainstream solutions to imbalanced binary classification fall into the following categories: surrogate loss functions [39, 59], cost-sensitive learning [17, 21, 103, 104], threshold selections [51], undersampling the majority class [49, 52, 73] or oversampling the minority class [11, 83, 95], and ensembling [24, 53, 77, 99]. Among these methods, only surrogate F_β-score loss functions [39, 46] demonstrated theoretical consistency with F_β-score under challenging scenarios. To the best of our knowledge, few methods were adopted for imbalanced data series classification [10, 11], and real-world IED detections.

On the other hand, the proposed iEDeaL framework employs a novel easy-to-use SaSu loss function as a surrogate for F_β-score. Compared to other F_β-score surrogates, SaSu enjoys straightforward formula and negative sampling-refined decision boundary.

Besides practical and theoretical considerations, we claim that iEDeaL has no conflict and can be combined with most existing methods of the other categories. For example, undersampling techniques [73] can be deployed in iEDeaL to sample non-IED instances.

[Surrogate F_β-score Loss Function] F_β-score is a common evaluation metric to avoid the majority bias in imbalanced binary classification [88]. However, its non-differentiability and non-decomposability prevent its direct usage as a loss function to train deep models. Non-differentiability refers to that F_β-score is derived from discrete counts, where gradients cannot be calculated and backpropagated. Non-decomposability refers to that F_β-score is a global metric, which cannot be combined from F_β-scores of mini-batches.

Theoretical machine learning studies have been conducted recently to tackle the problem of non-differentiability [19, 39, 44, 65, 66, 98] and non-decomposability [42, 60, 79, 96]. However, most of their F_β-consistency [59] analysis was built upon either linear models, which do not apply to deep models, or sophisticated assumptions, which are non-trivial for real-world verification. Moreover, the high time and space complexities (e.g., for fine-grain grid searches over cost-sensitive loss coefficients and decision thresholds [65, 79]) also prevent their application in training deep models.

On the contrary, we design iEDeaL and SaSu with more practical considerations than strong theoretical soundness. Built upon smooth functions [39] and a recent formula template [46], SaSu uses a differentiable and decomposable formula, rendering iEDeaL easy to adopt for scalable real-world IED detection. Its correctness is established by the fact is shares the same 1st-order stationary points [13] with F_β-score and empirically verified (against wBCE, ewBCE [17], Focal [50]) with 2 real IED datasets.

3 BACKGROUND

In this section, we briefly introduce the notations and definitions for surrogate F_β-score loss functions. Our results are built on several recent studies [21, 39, 46].

In the context of binary classification, entries in the confusion matrix could be rewritten into the following equation.

$$TP(\Theta) = \bar{c}_{11} = \sum_{i=1}^{n} y_i \hat{y}_i, \quad TN(\Theta) = \bar{c}_{00} = \sum_{i=1}^{n} (1-y_i)(1-\hat{y}_i)$$
$$FP(\Theta) = \bar{c}_{01} = \sum_{i=1}^{n} (1-y_i)\hat{y}_i, \quad FN(\Theta) = \bar{c}_{10} = \sum_{i=1}^{n} y_i(1-\hat{y}_i)$$

where n is the number of instances in the dataset, $y_i, \hat{y}_i \in \{0,1\}$ are the truth and predicted labels of the ith instance x_i, Θ is the parameter set (omitted when there is no ambiguity).

Precision, recall and F_β-score are confusion matrix-based evaluation metrics, defined in Equation 1.

$$P = \frac{\bar{c}_{11}}{\bar{c}_{11}+\bar{c}_{01}}, \quad R = \frac{\bar{c}_{11}}{\bar{c}_{11}+\bar{c}_{10}}, \quad F_\beta = \frac{(1+\beta^2)\cdot P \cdot R}{\beta^2 \cdot P + R} \tag{1}$$

where P is precision, R is recall.

We introduce $\bar{p}_{1*}$, $\bar{p}_{10}$, $\bar{p}_{01}$ to denote the sample proportions of positive instances, false negatives, and false positives as the following.

$$\bar{p}_{1*} = \frac{n_1}{n}, \quad \bar{p}_{10} = \frac{\bar{c}_{10}}{n_1}, \quad \bar{p}_{01} = \frac{\bar{c}_{01}}{n_0}$$

where n_1 and n_0 are the number of positive and negative instances, $n = n_1 + n_0$. We define the *imbalance ratio* using $r_{im} := n_0/n_1$. Based on the weak law of large numbers, the sample proportions converge to the true probabilities, i.e., $\bar{p}_{1*} \to p_{1*}$, $\bar{p}_{10} \to p_{10}$ and $\bar{p}_{01} \to p_{01}$, as the sample sizes n, n_1, and n_0 tend to infinity.

Using these sample proportions, the limit of F_β-score could be rewritten to the following equation.

$$\tilde{F}_\beta = \lim_{n\to\infty} F_\beta = \frac{(1+\beta^2)\cdot p_{1*}\cdot(1-p_{10})}{p_{1*}\cdot(\beta^2+1-p_{10}-p_{01})+p_{01}}$$

However, $\tilde{F}_\beta$ cannot be directly adopted as a loss function because p_{10} and p_{01} are non-decomposable and non-differentiable with respect to Θ. We do not require p_{1*} to be differentiable since it is dataset statistics without dependence on Θ.

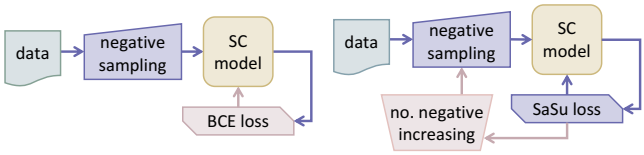
To work around the non-differentiability, a conventional tool is to approximate $\tilde{F}_\beta$ with p_{10} and p_{01}'s smooth functions. In SaSu, we adopt Equation 2 under the mild assumption $|\hat{y} - f(x;\Theta)| < 0.5$ [39].

$$\tilde{p}_{10} = \mathbb{E}[1-f(x;\Theta)|y=1], \quad \tilde{p}_{01} = \mathbb{E}[f(x;\Theta)|y=0] \tag{2}$$

where $f(x;\Theta) \in [0,1]$ is the classifier output given instance x.

4 THE iEDeaL FRAMEWORK

In this section, we present the proposed iEDeaL framework. The complete workflow is illustrated in Figure 3. Figure 3a shows the BCE-based SC model initialization, and Figure 3b shows the SaSu-based SC model training for F_β-score optimization. Given an EEG series collection, the negative sampling module draws a portion of negative instances to be combined with all positive instances in both procedures. The difference is that the number of negative instances for initialization is the same as positive instances (n_1), while the number of negative instances for training is gradually increased from the same number (n_1) to a pre-calculated maximal number



(a) Model initialization with the BCE loss function.

(b) Main procedure of the SaSu-based SC model training.

Figure 3: The iEDeaL model training framework. Purple indicates SaSu-guided operations. Pink indicates curriculum learning-based auxiliary training operations.

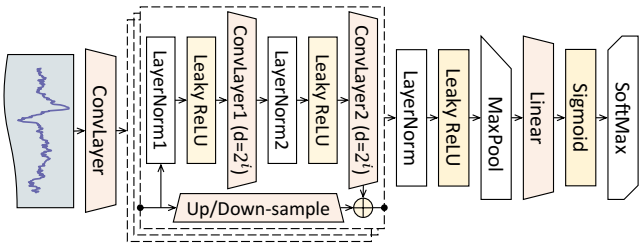


Figure 4: The SC architecture and the dilated full-preactivation ResBlock (in the dashed box).

($n_s - n_1$, where n_s is the size of the sampled training set). The module of increasing the number of negative instances determines the number of negative samples for each epoch.

We first present the SC architecture in Section 4.1. SC is adapted from the SEAnet architecture [89]. Its backbone structure is a unmasked full-preactivation ResNet [33]. It adopts the exponentially increasing dilations, which has been verified to be effective on raw data series [4]. We provide a formal definition of the SaSu loss function in Section 4.2. Its correctness is formally established based on the observation that SaSu shares (approximately) the same condition of first-order stationary points [13] as $\tilde{F}_\beta$-score, i.e., optimizing SaSu is equivalent to optimizing the F_β-score. We discuss in Section 4.3 the critical role negative sampling plays in deploying iEDeaL for real-world IED detection, i.e., in refining the decision boundaries and facilitating model convergence.

To further facilitate SC model convergence on SaSu, we propose a dynamically weighted regularization in Section 4.2.1, i.e., reducing gradients for well-classified instances [50], and two curriculum learning-based auxiliary training strategies in Section 4.4, i.e., BCE-based model initialization and increasing the number of negative samples, as in Figure 3. Benefiting from its symmetric structure, BCE serves as an easier target to initialize deep models for the harder SaSu loss function. Gradually increasing the number of negative samples helps by simply starting SC model training from the easier balanced mini-batches.

4.1 The SC Architecture

The SC architecture is illustrated in Figure 4. SC is a multi-layer residual CNN, stacking the full-preactivation ResBlocks (in the dashed box). The dilations for convolutional kernels increase exponentially with the depth of the ResBlocks. Compared with constant dilations, exponentially increasing dilations has been verified to effectively broaden the receptive fields for raw data series [4]. We determine the number of ResBlocks by making SC's receptive fields cover the whole series. The first ConvLayer targets to expand the channels and extract raw features. Other ConvLayers in ResBlocks take a kernel size of 3, following the conventions of ResNet [33]. The dimension of latent features is kept the same as the dimension of the input series for all ConvLayers.

4.2 The SaSu Loss Function

We now describe SaSu, a novel negative sampling-based surrogate F_β-score loss function. SaSu is defined in the following equation.

$$L_{SaSu}(f(x;\Theta), y) = -\, y \log f(x;\Theta) + (1-y)\log(\beta^2 \frac{\tilde{p}_{1*,s}}{\alpha(1-\tilde{p}_{1*,s})} + f(x;\Theta))$$

where $\alpha = (n - n_1)/(n_s - n_1) \geq 1$ is the number of all negative instances over the drawn negative samples. The sampling procedure only happens for the negative instances, i.e., we use all the positive instances in model training.

Given the above definition, it is easy to verify that L_{SaSu} is differentiable since L_{SaSu} is continuous on $f(x;\Theta) \in [0,1]$ and ∇L_{SaSu} exists as long as $\nabla f(x;\Theta)$ exists. L_{SaSu} is also intrinsically composable as it is defined at the instance level instead of the dataset level. Lemma 1 states that optimizing L_{SaSu} on a negative-sampled training set surrogates to optimize F_β-score on the whole training set in terms of approximately sharing the same condition of first-order stationary points [13]. (See full version for proof.)

Lemma 1. *Given an i.i.d. (independent and identically distributed) subset of negative instances $X_{0,s}$ from X_0, where $X_{0,s}$ and X_0 share the same underlying distribution $P(x|y=0,\Theta)$, L_{SaSu} on the negative-sampled training set $\{X_{0,s}, X_1\}$ approximately shares the same 1st-order stationary point conditions as F_β-score on the full training set.*

We note that Lemma 1's i.i.d. distribution assumption on negative samples is a simple and practical assumption in real-world scenarios. Similar assumptions have been made before while emphasizing slightly different perspectives [21, 98]. A *sufficient* number of *random* negative samples simply satisfies this assumption [17]. We present more details of iEDeaL's negative sampling in Section 4.3.

Another important observation from Lemma 1 is that $\tilde{p}_{1*,s}/\alpha(1-\tilde{p}_{1*,s}) = \tilde{p}_{1*}/(1-\tilde{p}_{1*})$. Intuitively, this observation proxies to train our model on a training set with a certain imbalance ratio, while targeting to deploy it on a real-world testing set with another different imbalance ratio. Similar insights have been made before [21] to theoretically support the undersampling techniques.

4.2.1 Regularizing SaSu. Although L_{SaSu} is capable of optimizing F_β-score and integrating negative sampling, its asymmetric formula introduces extra burdens to model convergence, compared to simple loss functions, e.g., BCE or MSE. We first discuss the problem of large gradients for well-classified negative instances in this section, and then tackle the asymmetric gradient problem in Section 4.4.

The equation of L_{SaSu}'s gradients for negative instances is in the following equation.

$$\nabla L_{SaSu}(f(x;\Theta), y=0) = \frac{1}{\beta^2 \frac{\tilde{p}_{1*,s}}{\alpha(1-\tilde{p}_{1*,s})} + f(x;\Theta)} \cdot \nabla f(x;\Theta)$$

For well-classified negative instances, i.e., $f(x;\Theta) \rightarrow 0, \nabla L_{SaSu}(f,0)$ has the largest coefficient $\rightarrow 1/(\beta^2 \bar{p}_{1*,s}/\alpha(1-\bar{p}_{1*,s}))$. This is an undesired property, introducing unnecessary turbulence around the convergence parameter hyper-regions.

To tackle this problem, we introduce modulating terms to regularize the gradients for well-classified instances, following the intuition behind the Focal loss function [50]. The regularized SaSu loss function is shown in the following equation:

$$L_{SaSu}(f(x;\Theta), y) = -(1-f(x))^{\gamma_1} y \log f(x;\Theta) + f(x)^{\gamma_0}(1-y)\log(\beta^2 \frac{\bar{p}_{1*,s}}{\alpha(1-\bar{p}_{1*,s})} + f(x;\Theta))$$

where $\gamma_1, \gamma_0 \in \mathbb{R}^+$ are hyperparameters. $f(x)$ denotes the output values (i.e., the predictions) of the SC models, detached from gradient propagation. For well-classified instances, the modulating terms go to 0 and the gradients are down-weighted to 0. For misclassified instances, the modulating terms are near 1 and the gradients are unaffected. Hence, the regularized SaSu loss function achieves more stable convergence points.

We use the regularized version of SaSu by default in the following sections. We fix $\gamma_1 = \gamma_0 = 1/2$ in our implementation to avoid suppressing the gradients to undesired small values, which empirically worked well across different cases in our experiments.

4.3 Negative Sampling in iEDeaL

We now describe how to draw negative samples satisfying the distribution assumption in Lemma 1. We propose two different domain-independent sampling strategies, i.e., to draw a *sufficient* number of *random* samples [17], or to draw a *smaller* number of data series index-based *representative* samples [89]. Choices between them depend on model complexity and computation resource, i.e., how many training samples can be supported.

The first sampling strategy is to draw a theoretically *sufficient* number of *random* negative samples. This idea naturally conforms to the law of large numbers. The key question is to derive the *minimal* sufficient number. One recent estimation [17] is $E(n_s - n_1) = (1-\beta^{n-n_1})/(1-\beta)$, where $\beta = (n-1)/n$, enlightened by the random covering problem [40]. We employ this strategy in our experiments due to its simplicity.

The second strategy focuses on drawing a *smaller* number of *selected* negative samples than the sufficient number of random samples. We propose to draw negative samples based on data series indexes [89, 100]. Similar negative instances are grouped adjacently in index-partitioned local groups (e.g., leaf nodes in tree indexes) [20, 63, 64, 67–71, 90], either in the input or latent feature space. By traversing the index structure, it is efficient to draw a small number of representative samples, covering all local groups, hence preserving the distribution. This strategy is promising to facilitate iEDeaL's deployment in extremely large-scale imbalanced scenarios [3], which we defer to future studies.

4.4 Curriculum Learning-based Auxiliary Training Strategies

SaSu introduces convergence challenges when compared to BCE/MSE due to its asymmetric structure, i.e., the gradients by mislabeling positive and negative instances are different. Moreover, the asymmetric numbers of positive and negative instances also introduce more gradient instability than balanced mini-batches. Thus, training SC models on SaSu in imbalanced datasets is more prone to gradient exploding/vanishing, resulting from both the asymmetric loss structure and the imbalanced mini-batches.

In order to better control SC model convergence on SaSu, we propose two curriculum learning-based auxiliary training strategies. The general idea of curriculum learning is to train models with increasing difficulty levels to benefit convergence [7]. The first strategy is to pretrain SC models with the BCE loss function for a few epochs, based on the observation that BCE and SaSu share the same gradient structure for mislabeling positive instances. The second strategy is to gradually increase the number of negative samples from the same number (n_1) as the positive instances to a pre-calculated maximal number ($n_s - n_1$), motivated by that models are harder to converge on more imbalanced datasets [66, 97]. By combining the two strategies with regularized SaSu into iEDeaL, the SC models enjoy more steady convergence to optimizing F_β-score on imbalanced IED datasets.

4.4.1 Initializing the SC model with BCE. The BCE loss function is formalized in the following equation:

$$L_{BCE}(f(x;\Theta), y) = -y \log f(x;\Theta) - (1-y)\log(1-f(x;\Theta))$$

Notably, L_{SaSu} and L_{BCE} share the same gradient formula for mislabeling positive instances. Hence, training on BCE for a few initial epochs guides deep models to the same parameter hyper-regions as SaSu for correctly classifying positive instances.

Although their negative gradient formula is different, generally they both help the model to correctly classify negative instances. As long as being stopped before falling into BCE's local optima and losing SaSu's momentum, BCE-based initialization should not damage the convergence to SaSu's local optima. This can be simply achieved by limiting the number of initial epochs and learning rates. Thus, training with L_{SaSu} after being initialized with L_{BCE} implies a better initial parameter state than random initialization.

4.4.2 Increasing the number of negative samples. Intuitively, balanced mini-batches are easier to provide stable gradients by avoiding internal covariate shifts [37], benefiting model convergence. Recent studies also confirm this intuition by finding that models are harder to converge to the optimal parameters on more imbalanced datasets [97]. Hence, we propose to gradually increase the number of negative samples from the same number (n_1) as the positive instances to a pre-calculated maximal number ($n_s - n_1$).

We implement this strategy using simple heuristics, i.e., to increase the number of negative samples at exponential rates. Specifically, we draw $\min(\lceil b^{\log_a e} \rceil n_1, n_s - n_1)$ negative samples at epoch e, where $a, b > 1$. For BCE-based model initialization, the number of negative samples is fixed to be n_1, i.e., by the balanced ratio. This implementation empirically worked well in our experiments with $a = \sqrt{2}, b = 2$, which generally ensures enough epochs left to fine-tune the model on the largest sample sets.

5 EXPERIMENTS

We present the experimental evaluation of the iEDeaL framework on two real-world IED datasets against other SOTA methods. In

summary, the results demonstrate that iEDeaL was 20% more accurate than other SOTA deep learning methods [23, 33], and 100x faster than existing SOTA spectrogram-based IED detection frameworks deployed in real neuroscience applications [28, 76].

[Setup] All deep models were trained using Nvidia Tesla V100 SXM2 (16G memory). Software environments were python/3.8.11, pytorch-gpu/py3/1.10.0 and cuda/11.0.

[Methods] We first evaluate iEDeaL against four SOTA deep learning methods of different flavors, ResNet-18 [33, 76], InceptionTime [23], TimeNet [56], and TST [101]. ResNet-18 and InceptionTime are convolutional models, TimeNet is a recurrent model, and TST is a Transformer model. ResNet-18 was originally designed for image classification [33], and further deployed in the AiED framework [76]. Since iEDeaL uses the raw series instead of the spectrogram in AiED, we evaluated iEDeaL against 1D ResNet-18 rather than 2D ResNet-18. InceptionTime [23] was widely considered as the SOTA data series classification algorithm [22]. We removed BatchNorm layers in ResNet-18 and InceptionTime since the highly imbalanced mini-batches damaged the stability of BatchNorm.

All models were trained up to 100 epochs using mini-batched Stochastic Gradient Descent (SGD). The batch size was set to 256. Learning rate was searched from {5*e*-3, 1*e*-2, 2*e*-2} and linearly decayed to 1*e*-5 [89]. Gradient clipping [102] was conducted for all methods with maximal norm = 1. Other hyperparameters were set to their default values. We used BCE in all other methods than iEDeaL, following their original designs. For iEDeaL, the number of BCE-based initialization epochs was set to be 3, $a = \sqrt{2}, b = 2$ for exponentially increasing the number of negative instances, and $\gamma_1 = \gamma_0 = 1/2$ for SaSu regularization. We evaluated these settings in sensitivity studies, as well as an index-based sampling method, SEAsam [89], against the default random sampling.

To further evaluate the effectiveness of SaSu in iEDeaL, we designed ablation experiments with SC models trained by other loss functions. These results could also be considered as evaluating SaSu as an imbalanced classification solution against other SOTA solutions, including ewBCE [17], Focal [50], and SF [46]. Specifically, the loss functions in SaSu's ablation experiments were constituted by: (i) *BCE*. (ii) *bBCE*, BCE on a balanced undersampled (randomly) training set. (iii) *wBCE*, the weighted BCE on the whole training set. Weights were set to be the inverse of positive/negative instance numbers, i.e., $1/n_1$ and $1/(n_s - n_1)$, and then normalized with sum = 2. (iv) *ewBCE*, the weighted BCE on a negative-sampled training set. The number of negative samples was set to $E(n_s - n_1)$ defined in Section 4.3 [17]. Weights were set similarly to wBCE. (v) *Focal*, the dynamically weighted BCE on the whole training set, as in Section 4.2.1 [50]. We set $\gamma = 2$ following the original designs [50]. (vi) *SF*, a surrogate F_β-score loss function [46] on the whole training set. Note that only SF and iEDeaL are capable to train deep models optimizing F_β-score with different β.

5.1 Datasets

Our experiments were conducted on two real-world EEG datasets, i.e., the ICM datasets [47] and the public TUEV datasets [72], covering a range of different imbalance levels ($0.97\% \leq p_1 \leq 24.6\%$). We present the dataset statistics in Table 1.

Table 1: Dataset statistics. n_1 denotes the number of positive instances, n_0 denotes the number of negative instances and p_1 denotes the percentage of positive instances.

	Train & Validation			Test		
Dataset	n_1	n_0	p_1 (%)	n_1	n_0	p_1 (%)
ICM1	10,923	71,529	13.2	2,185	14,306	13.2
ICM2	8,910	62,108	12.5	1,782	12,422	12.5
ICM3	6,001	80,698	6.92	1,201	16,140	6.93
ICM4	3,113	86,732	3.46	623	17,347	3.47
ICM5	2,635	82,599	3.09	528	16,520	3.10
ICM6	2,820	91,705	2.98	565	18,342	2.99
ICM7	2,304	86,138	2.61	461	17,228	2.61
ICM8	1,792	90,187	1.95	359	18,038	1.95
$TUEV_{3,3}$	18,083	65,949	21.5	7,242	22,179	24.6
$TUEV_{1,1}$	645	53,726	1.19	567	19,646	2.81
$TUEV_{1,3}$	645	65,949	0.97	567	22,179	2.49

[ICM] The ICM datasets are part of an ongoing study on interictal activity, consisting of data of eight epileptic patients from our database [47]. Patients were implanted with intracranial depth electrodes (Ad-Tech® Medical Instrument Corporation, Oak Creek, WI, USA), consisting of 4 to 8 macroelectrode contacts separated by 5 mm (1.3 mmØ). Trajectories, anatomical targets, number of electrodes, and number of electrode contacts, were determined according to the clinical practice and the epilepsy features of the patients. Implantation was performed at the Department of Neurosurgery of the Pitié-Salpêtrière Hospital using Leksell Model G stereotactic system (Elekta, Inc, Norcross, GA) or using the robotic assistant device ROSA (ROSA® Brain, Medtech, France). Signals from macroelectrodes were continuously recorded at 4 kHz, using a hardware high-pass filter at 0.01 Hz (Atlas Recording System, NeuraLynx, Tucson, AZ, USA). All patients gave written informed consent (project C11-16 conducted by INSERM and approved by the local ethics committee, CPP Paris VI).

For each patient, the first 24 hours were visually annotated for IEDs according to standard guidelines [34, 35], using software developed in-house (MUSE) [47]. Annotation and analyses were performed on the five deepest macroelectrode contacts, located within the amygdala-hippocampal complex, as part of ongoing research on IEDs in the medial-temporal lobe.

Signals were bipolar referenced to increase spatial resolution and reduce common noise (i.e., feature channel 1=raw channel 1 − raw channel 2) [48], filtered between 1 Hz to 50 Hz, and downsampled to 512 Hz. After being preprocessed, 1.5 s IED instances were first extracted. Non-IED instances were extracted for the first 24 hours using 1.5 s sliding windows, non-overlapping with the IED instances. Datasets for all patients were randomly split by 4 : 1 : 1 for the training, validation, and test sets.

[TUEV] The TUEV dataset is a subset of the public TUH EEG Corpus that contains annotations of EEG segments as one of six classes: spike and sharp wave (SPSW), generalized periodic epileptiform discharges (GPED), periodic lateralized epileptiform discharges (PLED), eye movement (EYEM), artifact (ARTF), and background (BCKG) The SPSW class corresponds exactly to the IED class in our context [32]. Following the TUEV usage conventions [28], we extract 3 subsets from the TUEV dataset: (i) $TUEV_{3,3}$, where SPSW, GPED, and PLED are positive classes, and EYEM, ARTF, and BCKG are negative classes; (ii) $TUEV_{1,1}$, where SPSW is the positive class

Table 2: The F_1-scores for different IED detection methods. Best result (higher is better) is marked in bold, second best is underlined. - indicates the result was not available (for HDL) or the model did not converge (for the rest).

Methods	ICM1	ICM2	ICM3	ICM4	ICM5	ICM6	ICM7	ICM8	$TUEV_{3,3}$	$TUEV_{1,1}$	$TUEV_{1,3}$	AvgRank
HDL [28]	-	-	-	-	-	-	-	-	0.53	0.36	0.27	5.41
TimeNet [56]	0.87	0.88	0.82	0.66	0.77	0.48	0.83	-	-	-	-	4.91
ResNet-18 (AiED [76])	0.87	0.90	0.89	0.73	0.81	0.56	0.79	0.72	**0.91**	0.57	-	3.59
InceptionTime [23]	0.87	0.91	0.86	0.69	0.79	0.57	0.81	0.70	0.89	0.51	0.43	3.45
TST [101]	0.91	0.94	0.91	0.85	0.90	0.73	0.86	0.87	0.31	**0.73**	0.12	2.36
iEDeaL	**0.95**	**0.96**	**0.96**	**0.90**	**0.94**	**0.80**	**0.91**	**0.90**	0.88	0.57	**0.49**	**1.27**

and BCKG is the negative class; (iii) $TUEV_{1,3}$, where SPSW is the positive class, and EYEM, ARTF, and BCKG are negative classes. Although the TUH EEG Corpus contains scalp EEG instead of SEEG, its annotations were conducted over single event occurrences rather than occurrence periods. Hence, it is suitable for evaluating iEDeaL.

The TUEV dataset was already split into training (including validation) and test sets. We further randomly split the given training set by 4 : 1 for training and validation. SOTA results on the TUEV datasets were obtained using the Hybrid Deep Learning (HDL) algorithm [28]. We derived binary classification results from the HDL confusion matrix by considering an instance being misclassified only if its prediction falls into an event of the opposite class.

Table 3: F_1-scores of SC models with different loss functions.

Dataset	bBCE	BCE	wBCE	ewBCE	Focal	SF	SaSu
ICM1	0.94	**0.95**	0.94	**0.95**	0.94	0.94	**0.95**
ICM2	**0.96**	**0.96**	0.95	**0.96**	0.95	**0.96**	**0.96**
ICM3	0.95	0.94	0.93	**0.96**	0.95	0.95	**0.96**
ICM4	0.80	0.90	0.90	0.90	0.90	**0.91**	0.90
ICM5	0.72	0.93	0.92	0.92	**0.94**	0.86	**0.94**
ICM6	0.59	0.77	0.76	0.76	0.74	0.72	**0.80**
ICM7	0.73	0.90	0.85	0.90	0.89	0.89	**0.91**
ICM8	0.58	0.88	0.63	0.81	0.88	0.84	**0.90**
$TUEV_{3,3}$	0.88	0.88	0.88	0.88	**0.90**	**0.90**	0.88
$TUEV_{1,1}$	0.48	0.47	0.46	0.54	0.49	-	**0.57**
$TUEV_{1,3}$	0.43	0.36	0.39	0.31	0.38	-	**0.49**
AvgRank	5.36	3.73	5.23	3.64	3.41	4.55	**2.09**

5.2 Results

In summary, the proposed iEDeaL framework provided higher F_1-scores than all other methods in 9 of 11 cases, achieving an average rank of 1.27/6. Moreover, iEDeaL was 100x faster than HDL and AiED, which were deployed in the real-world scenarios [28, 76], due to the fact that iEDeaL does not involve expensive feature extraction (e.g., spectrogram generation). In ablation experiments, we demonstrate the effectiveness of the SaSu loss function and the curriculum learning-based auxiliary training strategies in improving the F_1-score, achieving average ranks of 2.09/7 and 1.73/3, respectively. This verified SaSu to be an effective loss function for imbalanced classification, compared with other imbalanced classification solutions, including ewBCE [17] and Focal [50]. iEDeaL also provided higher F_2-scores than SC models trained using other loss functions, achieving an average rank of 1.86/7.

[F_1-scores for iEDeaL] The F_1-scores of all methods on each dataset are listed in Table 2. iEDeaL outperformed all other methods in 9 of 11 cases, achieving an average rank of 1.27/6. Specifically, iEDeaL has larger advancements than ResNet-18 and InceptionTime on more imbalanced ICM datasets ($\sim$ 20% on ICM4 to ICM8).

TST was the second best performer, behind iEDeaL, achieving an average rank of 2.36/6. TimeNet lagged behind other methods in 5 cases and did not converge in 4 cases, signaling the hardness of capturing EEG generative patterns due to their nonstationary morphology. ResNet-18 outperformed InceptionTime in 6 of 10 cases (excluding 1 tie) but did not converge in the most imbalanced $TUEV_{1,3}$, ending with an average rank of 3.59/6. The differences between iEDeaL, ResNet-18, and InceptionTime were the smallest on the most balanced $TUEV_{3,3}$, significantly outperforming HDL, which verified the effectiveness of raw EEG series-based deep learning methods over explicit feature extraction-based solutions.

[Ablation results for SaSu] We evaluated the effectiveness of the SaSu loss function in iEDeaL against SC models trained using other loss functions. The F_1-scores of all methods on different datasets are shown in Table 3.

The iEDeaL (i.e., the SaSu column) framework achieved the highest average rank of 2.09/7. It provided the highest F_1-scores, especially on more imbalanced datasets, e.g., ICM5 to ICM6, $TUEV_{1,1}$ and $TUEV_{1,3}$. SC+Focal reached 2nd place, with an average rank of 3.41/7. This verified the effectiveness of dynamical weighting in helping model convergence. SC+ewBCE function came 3rd place, with an average rank of 3.64/7. This observation followed the intuition that negative sampling benefits imbalanced IED detections. SC+BCE achieved an average rank of 3.73/7. We believe this benefited from BCE's easy convergence for SC models on the ICM dataset. SC+SF did not converge on $TUEV_{1,1}$ and $TUEV_{1,3}$, due to its imbalanced gradients for positive and negative instances. SC+bBCE had the lowest average rank of 5.36/7, demonstrating that training SC models on (randomly) balanced datasets generally cannot be deployed in real-world imbalanced scenarios.

Besides the F_1-scores, we also evaluated the F_2-scores. Detailed results were omitted due to the lack of space. F_2-score increases the importance of the precision and hence prefers solutions with fewer false alarms, which is generally a desired property in real-world IED detection applications. Notably, only iEDeaL and SC+SF can be trained targeting different β values in F_β-score. Other results were selected from the same set of trained models as in Table 3, using their F_2-scores on the validation set. The iEDeaL framework achieved the highest average rank of 1.86/7. SC+SF improved to 2nd place with an average rank of 3.82/7 for F_2-scores, from 4.55/7 for F_1-scores. SC+BCE and SC+Focal had follow-up performance than iEDeaL, benefited from their well-converged models. SC+bBCE kept the lowest average rank of 5.45/7.

To conclude, the F_1- and F_2-scores verified that SaSu in iEDeaL helped to train better SC models for imbalanced IED detection.

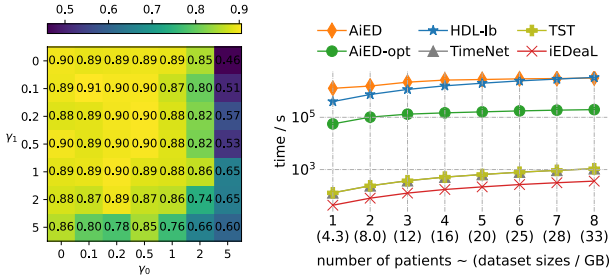
[Ablation results for the auxiliary training strategies] We evaluated the effectiveness of curriculum learning-based auxiliary

Table 4: F_1-scores for ablation experiments of the auxiliary training strategies.

Methods	ICM1	ICM2	ICM3	ICM4	ICM5	ICM6	ICM7	ICM8	$TUEV_{3,3}$	$TUEV_{1,1}$	$TUEV_{1,3}$	AvgRank
SC + SaSu	**0.95**	**0.97**	0.95	**0.91**	0.92	0.76	0.90	0.88	**0.89**	0.64	0.30	2.14
+ PreBCE	**0.95**	0.96	0.95	**0.91**	0.93	0.79	0.89	0.88	0.88	**0.65**	0.35	2.14
+ IncNeg = iEDeaL	**0.95**	0.96	**0.96**	0.90	**0.94**	**0.80**	**0.91**	**0.90**	0.88	0.57	**0.49**	**1.73**

Table 5: F_1-scores for the random negative sampling and index-based negative sampling (i.e., SEAsam).

	negative sample ratios $(n_s - n_1)/(n - n_1)$																
Methods	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	Avg
Random	0.56	0.61	0.56	**0.70**	0.61	0.62	0.54	**0.67**	**0.85**	**0.87**	0.87	0.88	**0.89**	0.87	**0.90**	0.89	0.74
SEAsam [89]	**0.70**	**0.65**	**0.64**	0.63	**0.67**	**0.70**	**0.58**	0.62	0.83	0.86	0.87	**0.89**	0.88	**0.88**	0.89	0.89	**0.76**



(a) F_1-scores for different γ_0 and γ_1 on the ICM8 dataset. **(b) Running time for detecting IEDs in the ICM dataset.**

Figure 5: (a) Sensitivity study for γ; (b) Running times.

training strategies in iEDeaL by ablation experiments. The F_1-scores on different datasets were reported in Table 4. PreBCE indicates model initialization with the BCE loss function. IncNeg indicates increasing the number of negative instances with more training epochs. Hence, iEDeaL=SC+SaSu+PreBCE+IncNeg.

As shown in Table 4, iEDeaL outperformed both SC+SaSu and SC+SaSu+PreBCE in 6/10 cases (excluding 1 tie), achieving an average rank of 1.73/3. SC+SaSu+PreBCE outperformed SC+SaSu in 4/7 cases, excluding 4 ties. Hence, we concluded that the auxiliary training strategies helped to train better SC models on SaSu for imbalanced IED detection.

[Sensitivity study for γ_1 and γ_0] We evaluated the sensitivity of γ_1 and γ_0 on the ICM8 dataset, and the results are shown in Figure 5a. The best performance region for γ_0 and γ_1 is within $[0.1, 0.5]$. Increasing γ_0 hurts model performance more heavily than increasing γ_1, confirming our main motivation to employ γ_0 and γ_1 for fixing the divergent gradients of negative instances.

[Sensitivity study for a and b] We evaluated the sensitivity of $a, b \in [\sqrt{2}, 4]$. The trained models achieved stable performance across different combinations, since enough epochs were left for fine-tuning on the largest sample sets. (Details omitted for brevity.)

[Negative sampling methods] We compared in Table 5 the random negative sampling strategy against the index-based negative sampling strategy, i.e., SEAsam [89]. When a small ratio (<10%) of negative samples were collected for training, SEAsam outperformed random sampling. With the increase of the sampling ratio, the performance difference between two methods diminishes.

[Detection time for iEDeaL] We compared the detection time of iEDeaL against the existing SOTA IED detection methods deployed in real neuroscience applications, including HDL [28] and AiED [75]. The reported numbers did not count the data preprocessing procedures described in Section 5.1, which were the same for all methods. AiED-opt is the best possible detection time of AiED, where we assume all IED instances pass the first template matching step while non-IED instances cannot pass at all. Since the code of HDL is not available, we estimated the HDL detection time using its spectrogram-based feature extraction (borrowed from AiED) time only, denoted by HDL-lb.

As shown in Figure 5b, iEDeaL was more than 100x faster than AiED-opt, AiED, and HDL-lb. The detection time of AiED and HDL was dominated by the spectrogram generation, showing the benefits of detecting on the raw series instead of on expensive explicit feature (e.g., spectrogram) extractions. Note also that iEDeaL was 3x faster than TST. Hence, we conclude that working directly on the raw EEG series, iEDeaL is not only an effective but also an extremely efficient method for real-world IED detection at scale.

6 CONCLUSIONS

In this paper, we propose the iEDeaL framework for real-world, highly imbalanced IED detection. When evaluated on real datasets, iEDeaL is more accurate and 2 orders of magnitude faster than the current SOTA methods, deployed in real applications, making it an important tool for clinicians and researchers. In our future work, we will study how to further improve the iEDeaL workflow, determining the minimal number of annotations required for both model training and transferring, in order to reduce time and increase objectivity in clinical research. Finally, we will study the use of iEDeaL and its interpretability [8] in other neuroscience applications that involve pattern detection in imbalanced datasets, including sporadic epileptiform discharges [34], benign sporadic sleep spikes and wicket spikes [55].

ACKNOWLEDGMENTS

Work supported by diiP, IdEx Université Paris Cité ANR-18-IDEX-0001, China Scholarship Council, HIPEAC 4, GENCI–IDRIS (Grant 2020-101471, 2021-101925, 2022-AD011012641R1), Investissements d'avenir ANR-10-IAIHU-06, Fondation de l'APHP pour la Recherche (Marie-Laure PLV Merchandising), and NVIDIA Corporation for the Titan Xp GPU donation used in this research. Part of this work was carried out on the CENIR-STIM core facility of ICM. We gratefully acknowledge Katia Lehongre for the technical support in data collection, and data management of the epilepsy database.

REFERENCES

[1] Andreas Antoniades, Loukianos Spyrou, David Martin-Lopez, Antonio Valentin, Gonzalo Alarcon, Saeid Sanei, and Clive Cheong Took. 2017. Detection of Interictal Discharges With Convolutional Neural Networks Using Discrete Ordered Multichannel Intracranial EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 12 (Dec. 2017), 2285–2294. https://doi.org/10.1109/TNSRE.2017.2755770

[2] Elham Bagheri, Justin Dauwels, Brian C. Dean, Chad G. Waters, M. Brandon Westover, and Jonathan J. Halford. 2017. Interictal Epileptiform Discharge Characteristics Underlying Expert Interrater Agreement. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology* 128, 10 (Oct. 2017), 1994–2005. https://doi.org/10.1016/j.clinph.2017.06.252

[3] Sara Bahaadini, Vahid Noroozi, Neda Rohani, Scott Coughlin, Michael Zevin, Joshua R. Smith, Vicky Kalogera, and Aggelos K. Katsaggelos. 2018. Machine learning for Gravity Spy: Glitch classification and dataset. *Information Sciences* (2018).

[4] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* (2018).

[5] Daniel T. Barkmeier, Aashit K. Shah, Danny Flanagan, Marie D. Atkinson, Rajeev Agarwal, Darren R. Fuerst, Kourosh Jafari-Khouzani, and Jeffrey A. Loeb. 2012. High Inter-Reviewer Variability of Spike Detection on Intracranial EEG Addressed by an Automated Multi-Channel Algorithm. *Clinical Neurophysiology* 123, 6 (June 2012), 1088–1095. https://doi.org/10.1016/j.clinph.2011.09.023

[6] Maxime O. Baud, Jonathan K. Kleen, Emily A. Mirro, Jason C. Andrechak, David King-Stephens, Edward F. Chang, and Vikram R. Rao. 2018. Multi-Day Rhythms Modulate Seizure Risk in Epilepsy. *Nature Communications* 9, 1 (Dec. 2018). https://doi.org/10.1038/s41467-017-02577-y

[7] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*.

[8] Paul Boniol, Mohammed Meftah, Emmanuel Remy, and Themis Palpanas. 2022. dCAM: Dimension-wise Class Activation Map for Explaining Multivariate Data Series Classification. In *SIGMOD*.

[9] Merritt W. Brown, Brenda E. Porter, Dennis J. Dlugos, Jeff Keating, Andrew B. Gardner, Phillip B. Storm, and Eric D. Marsh. 2007. Comparison of Novel Computer Detectors and Human Performance for Spike Detection in Intracranial EEG. *Clinical Neurophysiology* 118, 8 (Aug. 2007), 1744–1752. https://doi.org/10.1016/j.clinph.2007.04.017

[10] Hong Cao, Xiaoli Li, David Yew-Kwong Woon, and See-Kiong Ng. 2011. SPO: Structure Preserving Oversampling for Imbalanced Time Series Classification. In *ICDM*.

[11] Hong Cao, Xiaoli Li, David Yew-Kwong Woon, and See-Kiong Ng. 2013. Integrated Oversampling for Imbalanced Time Series Classification. *TKDE* (2013).

[12] Lei Cao, Wenbo Tao, Sungtae An, Jing Jin, Yizhou Yan, Xiaoyu Liu, Wendong Ge, Adam Sah, Leilani Battle, Jimeng Sun, Remco Chang, M. Brandon Westover, Samuel Madden, and Michael Stonebraker. 2019. Smile: A System to Support Machine Learning on EEG Data at Scale. *PVLDB* (2019).

[13] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. 2021. Lower bounds for finding stationary points II: first-order methods. *Mathematical Programming* (2021).

[14] Christopher J. L. Murray, Theo Vos, Rafael Lozano, Mohsen Naghavi, Abraham D. Flaxman, et al. 2012. Disability-Adjusted Life Years (DALYs) for 291 Diseases and Injuries in 21 Regions, 1990–2010: A Systematic Analysis for the Global Burden of Disease Study 2010. *The Lancet* 380, 9859 (2012), 2197–2223. https://doi.org/10.1016/S0140-6736(12)61689-4

[15] Erin C Conrad, Samuel B Tomlinson, Jeremy N Wong, Kelly F Oechsel, Russell T Shinohara, Brian Litt, Kathryn A Davis, and Eric D Marsh. 2020. Spatial Distribution of Interictal Spikes Fluctuates over Time and Localizes Seizure Onset. *Brain* 143, 2 (Feb. 2020), 554–569. https://doi.org/10.1093/brain/awz386

[16] Alexander C. Constantino, Nathaniel D. Sisterson, Naoir Zaher, Alexandra Urban, R. Mark Richardson, and Vasileios Kokkinos. 2021. Expert-Level Intracranial Electroencephalogram Ictal Pattern Detection by a Deep Learning Neural Network. *Frontiers in Neurology* 12 (May 2021), 603868. https://doi.org/10.3389/fneur.2021.603868

[17] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *CVPR*.

[18] Catarina da Silva Lourenço, Marleen C Tjepkema-Cloostermans, and Michel JAM van Putten. 2021. Machine learning for detection of interictal epileptiform discharges. *Clinical Neurophysiology* (2021).

[19] Krzysztof Dembczynski, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. 2011. An Exact Algorithm for F-Measure Maximization. In *NIPS*.

[20] Karima Echihabi, Panagiota Fatourou, Kostas Zoumpatianos, Themis Palpanas, and Houda Benbrahim. 2022. Hercules Against Data Series Similarity Search. *Proc. VLDB Endow.* 15, 10 (2022), 2005–2018.

[21] Charles Elkan. 2001. The foundations of cost-sensitive learning. In *IJCAI*.

[22] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *DMKD* (2019).

[23] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. 2020. InceptionTime: Finding AlexNet for time series classification. *DMKD* (2020).

[24] Everlandio R. Q. Fernandes, André C. P. L. F. de Carvalho, and Xin Yao. 2020. Ensemble of Classifiers Based on Multiobjective Genetic Sampling for Imbalanced Data. *TKDE* (2020).

[25] Kosuke Fukumori, Hoang Thien Thu Nguyen, Noboru Yoshida, and Toshihisa Tanaka. 2019. Fully Data-driven Convolutional Filters with Deep Learning Models for Epileptic Spike Detection. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2772–2776. https://doi.org/10.1109/ICASSP.2019.8682196

[26] Franz Fürbass, Mustafa Aykut Kural, Gerhard Gritsch, Manfred Hartmann, Tilmann Kluge, and Sándor Beniczky. 2020. An Artificial Intelligence-Based EEG Algorithm for Detection of Epileptiform EEG Discharges: Validation against the Diagnostic Gold Standard. *Clinical Neurophysiology* 131, 6 (June 2020), 1174–1179. https://doi.org/10.1016/j.clinph.2020.02.032

[27] David Geng, Ayham Alkhachroum, Manuel A. Melo Bicchi, Jonathan R. Jagid, Iahn Cajigas, and Zhe Sage Chen. 2021. Deep Learning for Robust Detection of Interictal Epileptiform Discharges. *Journal of Neural Engineering* 18, 5 (April 2021), 056015. https://doi.org/10.1088/1741-2552/abf28e

[28] Meysam Golmohammadi, Amir Hossein Harati Nejad Torbati, Silvia Lopez de Diego, Iyad Obeid, and Joseph Picone. 2019. Automatic Analysis of EEGs Using Big Data and Hybrid Deep Learning Architectures. *Frontiers in Human Neuroscience* (2019).

[29] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

[30] Jonathan J. Halford, Amir Arain, Giridhar P. Kalamangalam, Suzette M. LaRoche, Bonilha Leonardo, Maysaa Basha, Nabil J. Azar, Ekrem Kutluay, Gabriel U. Martz, Wolf J. Bethany, Chad G. Waters, and Brian C. Dean. 2017. Characteristics of EEG Interpreters Associated With Higher Interrater Agreement. *Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society* 34, 2 (March 2017), 168–173. https://doi.org/10.1097/WNP.0000000000000344

[31] Jonathan J. Halford, Robert J. Schalkoff, Jing Zhou, Selim R. Benbadis, William O. Tatum, Robert P. Turner, Saurabh R. Sinha, Nathan B. Fountain, Amir Arain, Paul B. Pritchard, Ekrem Kutluay, Gabriel Martz, Jonathan C. Edwards, Chad Waters, and Brian C. Dean. 2013. Standardized Database Development for EEG Epileptiform Transient Detection: EEGnet Scoring System and Machine Learning Analysis. *Journal of Neuroscience Methods* 212, 2 (Jan. 2013), 308–316. https://doi.org/10.1016/j.jneumeth.2012.11.005

[32] Amir Harati, Meysam Golmohammadi, Silvia Lopez, Iyad Obeid, and Joseph Picone. 2015. Improved EEG event classification using differential energy. In *SPMB*.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

[34] Lawrence J. Hirsch, Michael W. K. Fong, Markus Leitinger, Suzette M. LaRoche, Sandor Beniczky, Nicholas S. Abend, Jong Woo Lee, Courtney J. Wusthoff, Cecil D. Hahn, M. Brandon Westover, Elizabeth E. Gerard, Susan T. Herman, Hiba Arif Haider, Gamaleldin Osman, Andres Rodriguez-Ruiz, Carolina B. Maciel, Emily J. Gilmore, Andres Fernandez, Eric S. Rosenthal, Jan Claassen, Aatif M. Husain, Ji Yeoun Yoo, Elson L. So, Peter W. Kaplan, Marc R. Nuwer, Michel van Putten, Raoul Sutter, Frank W. Drislane, Eugen Trinka, and Nicolas Gaspard. 2021. American Clinical Neurophysiology Society's Standardized Critical Care EEG Terminology: 2021 Version. *Journal of Clinical Neurophysiology* 38, 1 (Jan. 2021), 1–29. https://doi.org/10.1097/WNP.0000000000000806

[35] L J Hirsch, S M LaRoche, N Gaspard, E Gerard, A Svoronos, S T Herman, R Mani, H Arif, N Jette, Y Minazad, J F Kerrigan, P Vespa, S Hantus, J Claassen, G B Young, E So, P W Kaplan, M R Nuwer, N B Fountain, and F W Drislane. 2013. American Clinical Neurophysiology Society's Standardized Critical Care EEG Terminology: 2012 Version. *Journal of Clinical Neurophysiology* 30, 1 (2013), 27.

[36] W. E. Hostetler, H. J. Doller, and R. W. Homan. 1992. Assessment of a Computer Program to Detect Epileptiform Spikes. *Electroencephalography and Clinical Neurophysiology* 83, 1 (July 1992), 1–11. https://doi.org/10.1016/0013-4694(92)90126-3

[37] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*.

[38] Radek Janca, Petr Jezdik, Roman Cmejla, Martin Tomasek, Gregory A. Worrell, Matt Stead, Joost Wagenaar, John G. R. Jefferys, Pavel Krsek, Vladimir Komarek, Premysl Jiruska, and Petr Marusic. 2015. Detection of Interictal Epileptiform Discharges Using Signal Envelope Distribution Modelling: Application to Epileptic and Non-Epileptic Intracranial Recordings. *Brain Topography* 28, 1 (Jan. 2015), 172–183. https://doi.org/10.1007/s10548-014-0379-1

[39] Martin Jansche. 2005. Maximum expected F-measure training of logistic regression models. In *HLT-EMNLP*.

[40] Svante Janson. 1986. Random coverings in several dimensions. *Acta Mathematica* (1986).

[41] Jin Jing, Aline Herlopian, Ioannis Karakis, Marcus Ng, Jonathan J. Halford, Alice Lam, Douglas Maus, Fonda Chan, Marjan Dolatshahi, Carlos F. Muniz, Catherine Chu, Valeria Sacca, Jay Pathmanathan, WenDong Ge, Haoqi Sun, Justin Dauwels, Andrew J. Cole, Daniel B. Hoch, Sydney S. Cash, and M. Brandon Westover. 2020. Interrater Reliability of Experts in Identifying Interictal Epileptiform Discharges in Electroencephalograms. *JAMA Neurology* 77, 1 (Jan. 2020), 49–57. https://doi.org/10.1001/jamaneurol.2019.3531

[42] Purushottam Kar, Harikrishna Narasimhan, and Prateek Jain. 2014. Online and Stochastic Gradient Methods for Non-decomposable Loss Functions. In *NIPS*.

[43] Philippa J Karoly, Dean R Freestone, Ray Boston, David B Grayden, David Himes, Kent Leyde, Udaya Seneviratne, Samuel Berkovic, Terence O'Brien, and Mark J Cook. 2016. Interictal spikes and epileptic seizures: their relationship and underlying rhythmicity. *Brain* 139, 4 (2016), 1066–1078.

[44] Oluwasanmi Koyejo, Nagarajan Natarajan, Pradeep Ravikumar, and Inderjit S. Dhillon. 2014. Consistent Binary Classification with Generalized Performance Metrics. In *NIPS*.

[45] Balu Krishnan, Ioannis Vlachos, Aaron Faith, Steven Mullane, Korwyn Williams, Andreas Alexopoulos, and Leonidas Iasemidis. 2014. A Novel Spatiotemporal Analysis of Peri-Ictal Spiking to Probe the Relation of Spikes and Seizures in Epilepsy. *Annals of biomedical engineering* 42, 8 (Aug. 2014), 1606. https://doi.org/10.1007/s10439-014-1004-x

[46] Namgil Lee, Heejung Yang, and Hojin Yoo. 2021. A surrogate loss function for optimization of F_β score in binary classification with imbalanced data. *arXiv* 2104.01459 (2021).

[47] Katia Lehongre, Virginie Lambrecq, Stephen Whitmarsh, Valerio Frazzini, Louis Cousyn, Daniel Soleil, Sara Fernandez-Vidal, Bertrand Mathon, Marion Houot, Jean-Didier Lemarechal, Stéphane Clemenceau, Dominique Hasboun, Claude Adam, and Vincent Navarro. 2022. Long-Term Deep Intracerebral Microelectrode Recordings in Patients with Drug-Resistant Epilepsy: Proposed Guidelines Based on 10-Year Experience. *NeuroImage* (March 2022), 119116. https://doi.org/10.1016/j.neuroimage.2022.119116

[48] Guangye Li, Shize Jiang, Sivylla E. Paraskevopoulou, Meng Wang, Yang Xu, Zehan Wu, Liang Chen, Dingguo Zhang, and Gerwin Schalk. 2018. Optimal Referencing for Stereo-Electroencephalographic (SEEG) Recordings. *NeuroImage* 183 (Dec. 2018), 327–335. https://doi.org/10.1016/j.neuroimage.2018.08.020

[49] Lusi Li, Haibo He, and Jie Li. 2020. Entropy-based Sampling Approaches for Multi-Class Imbalanced Problems. *TKDE* (2020).

[50] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *ICCV*.

[51] Zachary Chase Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. 2014. Optimal Thresholding of Classifiers to Maximize F1 Measure. In *ECML-PKDD*.

[52] Chien-Liang Liu and Po-Yen Hsieh. 2020. Model-Based Synthetic Sampling for Imbalanced Data. *TKDE* (2020).

[53] Zhining Liu, Wei Cao, Zhifeng Gao, Jiang Bian, Hechang Chen, Yi Chang, and Tie-Yan Liu. 2020. Self-paced Ensemble for Highly Imbalanced Massive Data Classification. In *ICDE*.

[54] Catarina Lourenço, Marleen C. Tjepkema-Cloostermans, Luís F. Teixeira, and Michel J. A. M. van Putten. 2020. Deep Learning for Interictal Epileptiform Discharge Detection from Scalp EEG Recordings. In *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019 (IFMBE Proceedings)*, Jorge Henriques, Nuno Neves, and Paulo de Carvalho (Eds.). Springer International Publishing, Cham, 1984–1997. https://doi.org/10.1007/978-3-030-31635-8_237

[55] Greta Macorig, Arielle Crespel, Annacarmen Nilo, Ngoc Phuong Loc Tang, Mariarosaria Valente, Gian Luigi Gigli, and Philippe Gélisse. 2021. Benign EEG Variants in the Sleep–Wake Cycle: A Prospective Observational Study Using the 10–20 System and Additional Electrodes. *Neurophysiologie Clinique* 51, 3 (June 2021), 233–242. https://doi.org/10.1016/j.neucli.2021.03.006

[56] Pankaj Malhotra, Vishnu TV, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. 2017. TimeNet: Pre-trained deep recurrent neural network for time series classification. In *ESANN*.

[57] A. V. Medvedev, G. I. Agoureeva, and A. M. Murro. 2019. A Long Short-Term Memory Neural Network for the Detection of Epileptiform Spikes and High Frequency Oscillations. *Scientific Reports* 9, 1 (Dec. 2019), 19374. https://doi.org/10.1038/s41598-019-55861-w

[58] P. Megevand, L. Spinelli, M. Genetti, V. Brodbeck, S. Momjian, K. Schaller, C. M. Michel, S. Vulliemoz, and M. Seeck. 2014. Electric Source Imaging of Interictal Activity Accurately Localises the Seizure Onset Zone. *Journal of Neurology, Neurosurgery & Psychiatry* 85, 1 (Jan. 2014), 38–43. https://doi.org/10.1136/jnnp-2013-305515

[59] Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. 2013. On the statistical consistency of algorithms for binary classification under class imbalance. In *ICML*.

[60] Harikrishna Narasimhan, Purushottam Kar, and Prateek Jain. 2015. Optimizing Non-decomposable Performance Measures: A Tale of Two Classes. In *ICML*.

[61] Antoine Nonclercq, Martine Foulon, Denis Verheulpen, Cathy De Cock, Marga Buzatu, Pierre Mathys, and Patrick Van Bogaert. 2012. Cluster-Based Spike Detection Algorithm Adapts to Interpatient and Intrapatient Variation in Spike Morphology. *Journal of Neuroscience Methods* 210, 2 (Sept. 2012), 259–265. https://doi.org/10.1016/j.jneumeth.2012.07.015

[62] Andre Palmini, Antonio Gambardella, Frederick Andermann, Francois Dubeau, Jaderson C. da Costa, Andre Olivier, Donatella Tampieri, Pierre Gloor, Felipe Quesney, Eva Andermann, Eduardo Paglioli, Eliseu Paglioli-Neto, Ligia Coutinho Andermann, Richard Leblanc, and Hyoung-Ihl Kim. 1995. Intrinsic Epileptogenicity of Human Dysplastic Cortex as Suggested by Corticography and Surgical Results. *Annals of Neurology* 37, 4 (April 1995), 476–487. https://doi.org/10.1002/ana.410370410

[63] Themis Palpanas. 2015. Data Series Management: The Road to Big Sequence Analytics. *SIGMOD Record* (2015).

[64] Themis Palpanas. 2020. Evolution of a Data Series Index - The iSAX Family of Data Series Indexes. In *Communications in Computer and Information Science (CCIS)*, Vol. 1197.

[65] Shameem Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. 2014. Optimizing F-Measures by Cost-Sensitive Classification. In *NIPS*.

[66] Joan Pastor-Pellicer, Francisco Zamora-Martínez, Salvador España Boquera, and María José Castro Bleda. 2013. F-Measure as the Error Function to Train Neural Networks. In *IWANN*.

[67] Botao Peng, Panagiota Fatourou, and Themis Palpanas. 2018. ParIS: The Next Destination for Fast Data Series Indexing and Query Answering. *IEEE BigData* (2018).

[68] Botao Peng, Panagiota Fatourou, and Themis Palpanas. 2020. MESSI: In-Memory Data Series Indexing. In *ICDE*.

[69] Botao Peng, Panagiota Fatourou, and Themis Palpanas. 2020. ParIS+: Data Series Indexing on Multi-core Architectures. *TKDE* (2020).

[70] Botao Peng, Panagiota Fatourou, and Themis Palpanas. 2021. Fast data series indexing for in-memory data. *VLDB J.* 30, 6 (2021).

[71] Botao Peng, Panagiota Fatourou, and Themis Palpanas. 2021. SING: Sequence Indexing Using GPUs. In *Proceedings of the International Conference on Data Engineering (ICDE)*.

[72] Joseph Picone. 2015. Temple University EEG Corpus. https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml. Accessed: 2022-11-18.

[73] Angélica Guzmán Ponce, José Salvador Sánchez, Rosa Maria Valdovinos, and José Raymundo Marcial-Romero. 2021. DBIG-US: A two-stage under-sampling algorithm to face the class imbalance problem. *ESWA* 168 (2021), 114301.

[74] Timothée Proix, Wilson Truccolo, Marc G. Leguia, Thomas K. Tcheng, David King-Stephens, Vikram R. Rao, and Maxime O. Baud. 2021. Forecasting Seizure Risk in Adults with Focal Epilepsy: A Development and Validation Study. *The Lancet. Neurology* 20, 2 (Feb. 2021), 127–135. https://doi.org/10.1016/S1474-4422(20)30396-3

[75] Robert J Quon, Stephen Meisenhelter, Edward J Camp, Markus E Testorf, Yinchen Song, Qingyuan Song, George W Culler, Payam Moein, and Barbara C Jobst. 2021. Dartmouth ECoG Lab Automated Spike Detector. https://github.com/ecoglab/aied. Accessed: 2022-11-18.

[76] Robert J Quon, Stephen Meisenhelter, Edward J Camp, Markus E Testorf, Yinchen Song, Qingyuan Song, George W Culler, Payam Moein, and Barbara C Jobst. 2022. AiED: Artificial intelligence for the detection of intracranial interictal epileptiform discharges. *Clinical Neurophysiology* (2022).

[77] Roozbeh Razavi-Far, Maryam Farajzadeh-Zanjani, Boyu Wang, Mehrdad Saif, and Shiladitya Chakrabarti. 2021. Imputation-Based Ensemble Techniques for Class Imbalance Learning. *TKDE* (2021).

[78] E. E. M. Reus, F. M. E. Cox, J. G. van Dijk, and G. H. Visser. 2022. Automated Spike Detection: Which Software Package? *Seizure* 95 (Feb. 2022), 33–37. https://doi.org/10.1016/j.seizure.2021.12.012

[79] Amartya Sanyal, Pawan Kumar, Purushottam Kar, Sanjay Chawla, and Fabrizio Sebastiani. 2018. Optimizing non-decomposable measures with deep networks. *Machine Learning* (2018).

[80] Mark L. Scheuer, Anto Bagic, and Scott B. Wilson. 2017. Spike Detection: Inter-reader Agreement and a Statistical Turing Test on a Large Data Set. *Clinical Neurophysiology* 128, 1 (Jan. 2017), 243–250. https://doi.org/10.1016/j.clinph.2016.11.005

[81] Kevin J Staley and F Edward Dudek. 2006. Interictal Spikes and Epileptogenesis. *Epilepsy Currents* 6, 6 (Nov. 2006), 199–202. https://doi.org/10.1111/j.1535-7511.2006.00145.x

[82] Hans Stroink, Robbert-Jan Schimsheimer, Al W. de Weerd, Ada T. Geerts, Willem F. Arts, Els A. Peeters, Oebele F. Brouwer, A. Boudewijn Peters, and Cees A. van Donselaar. 2006. Interobserver Reliability of Visual Interpretation of Electroencephalograms in Children with Newly Diagnosed Seizures. *Developmental Medicine and Child Neurology* 48, 5 (May 2006), 374–377. https://doi.org/10.1017/S0012162206000806

[83] Yi Sun, Lijun Cai, Bo Liao, and Wen Zhu. 2022. Minority Sub-Region Estimation-Based Oversampling for Imbalance Learning. *TKDE* (2022).

[84] Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha. 2019. Data Augmentation Using GANs. *arXiv:1904.09135 [cs, stat]* (April 2019). arXiv:1904.09135 [cs, stat]

[85] Marleen C. Tjepkema-Cloostermans, Rafael C.V. de Carvalho, and Michel J.A.M. van Putten. 2018. Deep Learning for Detection of Focal Epileptiform Discharges

from Scalp EEG Recordings. *Clinical Neurophysiology* 129, 10 (Oct. 2018), 2191–2196. https://doi.org/10.1016/j.clinph.2018.06.024

[86] C. A. van Donselaar, R.-J. Schimsheimer, A. T. Geerts, and A. C. Declerck. 1992. Value of the Electroencephalogram in Adult Patients With Untreated Idiopathic First Seizures. *Archives of Neurology* 49, 3 (March 1992), 231–237. https://doi.org/10.1001/archneur.1992.00530270045017

[87] Pieter van Mierlo, Gregor Strobbe, Vincent Keereman, Gwénael Birot, Stefanie Gadeyne, Markus Gschwind, Evelien Carrette, Alfred Meurs, Dirk Van Roost, Kristl Vonck, Margitta Seeck, Serge Vulliémoz, and Paul Boon. 2017. Automated Long-Term EEG Analysis to Localize the Epileptogenic Zone. *Epilepsia Open* 2, 3 (2017), 322–333. https://doi.org/10.1002/epi4.12066

[88] Cornelis Joost Van Rijsbergen. 1974. Foundation of evaluation. *Journal of documentation* (1974).

[89] Qitong Wang and Themis Palpanas. 2021. Deep Learning Embeddings for Data Series Similarity Search. In *KDD*.

[90] Zeyu Wang, Qitong Wang, Peng Wang, Themis Palpanas, and Wei Wang. 2023. Dumpy: A Compact and Adaptive Index for Large Data Series Collections. In *SIGMOD*.

[91] W.R.S. Webber, B. Litt, R.P. Lesser, R.S. Fisher, and I. Bankman. 1993. Automatic EEG Spike Detection: What Should the Computer Imitate? *Electroencephalography and Clinical Neurophysiology* 87, 6 (Dec. 1993), 364–373. https://doi.org/10.1016/0013-4694(93)90149-P

[92] S.B. Wilson, R.N. Harner, F.H. Duffy, B.R. Tharp, M.R. Nuwer, and M.R. Sperling. 1996. Spike Detection. I. Correlation and Reliability of Human Experts. *Electroencephalography and Clinical Neurophysiology* 98, 3 (March 1996), 186–198. https://doi.org/10.1016/0013-4694(95)00221-9

[93] Scott B. Wilson and Ronald Emerson. 2002. Spike Detection: A Review and Comparison of Algorithms. *Clinical Neurophysiology* 113, 12 (Dec. 2002), 1873–1881. https://doi.org/10.1016/S1388-2457(02)00297-3

[94] Gregory A. Worrell, Terrence D. Lagerlund, and Jeffrey R. Buchhalter. 2002. Role and Limitations of Routine and Ambulatory Scalp Electroencephalography in Diagnosing and Managing Seizures. *Mayo Clinic Proceedings* 77, 9 (Sept. 2002), 991–998. https://doi.org/10.4065/77.9.991

[95] Yuxi Xie, Min Qiu, Haibo Zhang, Lizhi Peng, and Zhenxiang Chen. 2022. Gaussian Distribution Based Oversampling for Imbalanced Data Classification. *TKDE* (2022).

[96] Bowei Yan, Oluwasanmi Koyejo, Kai Zhong, and Pradeep Ravikumar. 2018. Binary Classification with Karmic, Threshold-Quasi-Concave Metrics. In *ICML*.

[97] Yan Yan, Tianbao Yang, Yi Yang, and Jianhui Chen. 2017. A Framework of Online Learning with Imbalanced Streaming Data. In *AAAI*.

[98] Nan Ye, Kian Ming Chai, Wee Sun Lee, and Hai Leong Chieu. 2012. Optimizing F-measures: a tale of two approaches. In *ICML*.

[99] Jian Yin, Chunjing Gan, Kaiqi Zhao, Xuan Lin, Zhe Quan, and Zhi-Jie Wang. 2020. A Novel Model for Imbalanced Data Classification. In *AAAI*.

[100] Manzil Zaheer, Satwik Kottur, Amr Ahmed, José Moura, and Alex Smola. 2017. Canopy Fast Sampling with Cover Trees. In *ICML*.

[101] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. 2021. A Transformer-based Framework for Multivariate Time Series Representation Learning. In *KDD*.

[102] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. 2020. Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity. In *ICLR*.

[103] Peilin Zhao, Yifan Zhang, Min Wu, Steven C. H. Hoi, Mingkui Tan, and Junzhou Huang. 2019. Adaptive Cost-Sensitive Online Classification. *TKDE* (2019).

[104] Zhi-Hua Zhou and Xu-Ying Liu. 2006. Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem. *TKDE* (2006).

[105] Maeike Zijlmans, Willemiek Zweiphenning, and Nicole van Klink. 2019. Changing Concepts in Presurgical Assessment for Epilepsy Surgery. *Nature Reviews Neurology* 15, 10 (Oct. 2019), 594–606. https://doi.org/10.1038/s41582-019-0224-y