Subject-Independent Detection of Ictal EEG in Pediatric Intractable Epilepsy

A Leave-Subjects-Out Machine-Learning Analysis of the CHB-MIT Scalp-EEG Database

July 11, 2026

Abstract

Automated detection of seizures from scalp electroencephalography (EEG) could reduce the profound under-reporting that undermines the clinical management of children with drug-resistant epilepsy, but performance figures in the literature are frequently inflated by patient-specific training and by evaluation schemes that allow information to leak between correlated recordings of the same child. This report develops and rigorously evaluates a *subject-independent* classifier that labels fixed-length windows of pediatric scalp EEG as ictal (seizure) or interictal. Using the Children's Hospital Boston–MIT (CHB-MIT) Scalp EEG Database, we analysed six children (30 clinician-annotated seizures) over an 18-channel bipolar montage sampled at 256 Hz. Continuous recordings were filtered (0.5 Hz to 45 Hz band-pass, 60 Hz notch) and segmented into 4 s windows with 50 % overlap, from which 180 features (variance, line length, higher-order moments, five relative spectral band powers, and spectral entropy per channel) were extracted. A random-forest classifier was trained and evaluated under strict six-fold Leave-Subjects-Out Cross-Validation (LOSO-CV), such that no window from a held-out child ever appeared in training. Severe class imbalance (2.75 % ictal prevalence) was addressed with majority-class downsampling to a 10 : 1 ratio and balanced class weighting on the training folds only, leaving the natural test-fold prevalence intact. Pooled across all held-out windows the model achieved an AUROC of 0.979, an AUPRC of 0.650 (a 23.7× lift over the 0.027 prevalence baseline), and a sensitivity of 89.4 % at a fixed 5 % false-positive rate. Macro-averaged across children the corresponding values were 0.987, 0.816, and 93.9 %. Feature-importance analysis showed that time-domain amplitude and waveform-complexity measures—variance and line length—accounted for roughly 86 % of total predictive importance, concentrated on central-midline (CZ-PZ, FZ-CZ) and right-temporal (T8-P8) derivations. The results demonstrate that a compact, interpretable feature set generalises across unseen pediatric subjects, while the per-subject variability (AUPRC 0.61 to 0.98) quantifies the residual gap that any clinically deployable detector must still close. Per-window predictions and code accompany this report.

Executive Summary

Objective. Build and honestly evaluate a machine-learning model that distinguishes seizure (ictal) from non-seizure (interictal) windows of pediatric scalp EEG, using a subject-independent protocol that reflects how such a tool would face a *new* patient in the clinic.

Data. Six children (chb01, chb02, chb03, chb05, chb07, chb08) from the CHB-MIT Scalp EEG Database; 18 bipolar channels at 256 Hz; 30 clinician-annotated seizures totalling 47.0 min; 48 EDF recordings yielding 49 959 labelled windows (1372 ictal, 2.75 %).

Method. 4 s windows, 50 % overlap; 180 handcrafted features/window; random forest (400 trees); six-fold Leave-Subjects-Out Cross-Validation; 10 : 1 downsampling plus balanced weighting on training folds only; leakage-free standardisation.

Headline results. Pooled AUROC 0.979; AUPRC 0.650 vs. a 0.027 baseline (23.7×); sensitivity 89.4 % at a fixed 5 % false-positive rate. Every held-out child exceeded AUROC 0.97.

Interpretation. Variance and line length dominate (∼86 % of importance); the most informative derivations are CZ-PZ, FZ-CZ and T8-P8. Performance is strong and generalises across unseen children, but AUPRC varies substantially by subject (0.61 to 0.98), underscoring that subject-to-subject heterogeneity—not average accuracy—is the principal obstacle to clinical translation.

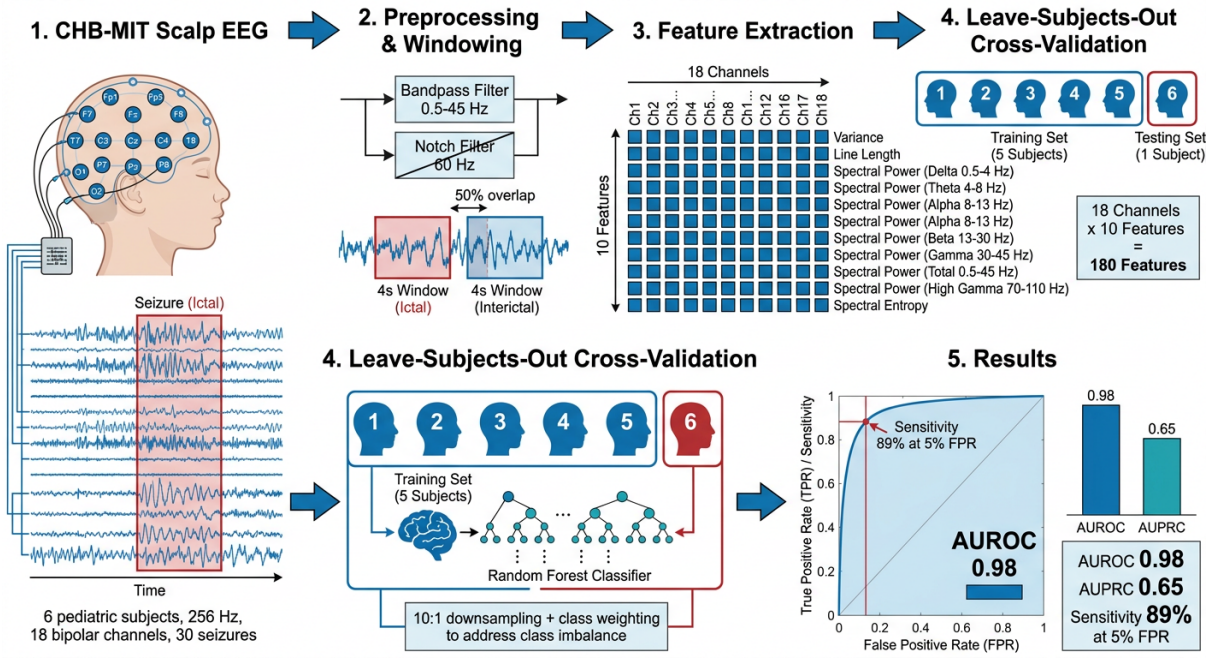

Figure 1. Graphical abstract. End-to-end workflow of the study: acquisition of pediatric scalp EEG from the CHB-MIT database (six subjects, 256 Hz, 18 bipolar channels), preprocessing and overlapping-window segmentation, extraction of 180 time- and frequency-domain features, subject-independent Leave-Subjects-Out cross-validation with imbalance-aware random-forest training, and evaluation with imbalance-robust metrics.

1 Introduction

1.1 Clinical burden of pediatric drug-resistant epilepsy

Epilepsy is among the most common serious neurological disorders, and the Global Burden of Disease programme estimated that tens of millions of people live with active epilepsy worldwide, with a disproportionate share of the burden borne by children and by low-resource settings [1, 2].

Approximately one third of patients continue to have seizures despite appropriate pharmacotherapy; the International League Against Epilepsy (ILAE) formalised this group as having drug-resistant epilepsy—defined as the failure of two adequately chosen, tolerated antiseizure medication schedules to achieve sustained seizure freedom [3]. The clinical consequences of failing to identify and act on refractoriness early are well documented [4, 5]. In children, persistent uncontrolled seizures carry not only developmental and psychosocial costs but also a measurable mortality risk through sudden unexpected death in epilepsy (SUDEP), for which generalised tonic–clonic seizures and poor seizure control are the strongest risk factors [6, 7].

1.2 The problem of seizure quantification

A recurring obstacle across all of these clinical decisions is that seizures are counted unreliably. Patient- and caregiver-reported seizure diaries systematically under-count events—particularly nocturnal, focal, and impaired-awareness seizures that the patient does not remember—so that objective seizure burden is often badly mis-estimated [8]. Continuous video-EEG monitoring remains the reference standard, but manual review of multi-day, multi-channel recordings is labour-intensive and does not scale to the ambulatory, long-term monitoring that would be most clinically useful. This gap has motivated two decades of work on automated seizure detection and on wearable devices, now the subject of formal ILAE / IFCN clinical practice guidance [9, 10]. Reviews and meta-analyses of such systems consistently identify the same tension: high sensitivity is achievable, but the *false-alarm rate* that patients and clinicians will tolerate is very low, and it is here that most systems still fall short [11, 12].

1.3 The CHB-MIT database and its evaluation pitfalls

The CHB-MIT Scalp EEG Database, collected at Children's Hospital Boston and released through PhysioNet, is the canonical public benchmark for pediatric scalp-EEG seizure detection [13–15]. It comprises continuous recordings from children with intractable epilepsy, digitised at 256 Hz in the international 10–20 system, with clinician-marked seizure onset and offset times. The foundational work of Shoeb and Guttag established that *patient-specific* classifiers— trained and tested on the same individual—can detect the large majority of seizures with short latency and few false detections [15, 16]. However, a patient-specific detector must be tuned on labelled seizures from the very patient it will serve, which is precisely the data one does not have when a child first presents. Subsequent reviews have repeatedly cautioned that the headline accuracies reported on CHB-MIT are frequently inflated by two methodological choices: patient-specific training, and cross-validation splits that mix windows from the same recording—or the same seizure—across training and test sets [17, 18]. Because seizures cluster within subjects and evolve slowly within an event, adjacent or overlapping windows are highly correlated; allowing them to straddle the train/test boundary leaks subject- and event-specific information and yields optimistic performance that does not transfer to new patients.

1.4 Objectives and contributions

This report addresses that gap directly by building and evaluating a *subject-independent* ictal-vs-interictal window classifier under a strict Leave-Subjects-Out protocol. Our contributions are fourfold. First, we assemble a leakage-controlled pipeline in which every window of a held-out child is excluded from training, so that reported performance reflects generalisation to genuinely unseen subjects. Second, we handle the severe (2.75 %) class imbalance explicitly and transparently, applying resampling and class weighting to the training folds only while preserving the natural clinical prevalence in the test folds. Third, we evaluate with metrics that are robust to imbalance—AUROC, AUPRC referenced against prevalence, and sensitivity at

a *fixed* false-positive rate—and we report the full distribution of performance across held-out children rather than a single pooled number. Fourth, we interpret the model through cross-fold feature importance, linking predictive signal to specific waveform properties and scalp derivations. Per-window predictions and the complete code are released with this report to support reproducibility.

2 Data and Methods

2.1 Database and subject selection

All data were drawn from the CHB-MIT Scalp EEG Database (version 1.0.0) hosted on PhysioNet [13, 19]. We restricted the analysis to a fixed subset of six subjects—`chb01`, `chb02`, `chb03`, `chb05`, `chb07`, and `chb08`—selected because they share a consistent 18-channel bipolar montage and together provide a range of seizure counts, durations, and ages (Table 1). Across these children the database provides 30 clinician-annotated seizures spanning 47.0 min of ictal time, with per-seizure durations ranging from 9 s to 264 s. Recordings are stored in European Data Format (EDF), with seizure onset/offset times supplied in accompanying annotation files; these annotations constitute the ground-truth labels for this study.

Table 1. Subjects analysed from the CHB-MIT database. Age and sex are as recorded in the database documentation; seizure counts and durations are computed from the clinician annotation files used as ground truth. "Ictal windows" and "Interictal windows" are the labelled 4 s windows entering the model (see Section 2.4).

Subject	Sex	Age (y)	Seizures	Ictal time (s)	Ictal windows	Interictal windows
chb01	F	11.0	7	442	212	5,397
chb02	M	11.0	3	172	82	5,397
chb03	F	14.0	7	402	192	5,397
chb05	F	7.0	5	558	273	5,397
chb07	F	14.5	3	325	159	21,602
chb08	M	3.5	5	919	454	5,397
Total	—	—	**30**	**2,818**	**1,372**	**48,587**

2.2 Channel montage

We used the 18-channel longitudinal bipolar ("double banana") montage common to these recordings, comprising four parasagittal–temporal chains and a midline chain: the left temporal chain (FP1-F7, F7-T7, T7-P7, P7-O1), the left parasagittal chain (FP1-F3, F3-C3, C3-P3, P3-O1), the right parasagittal chain (FP2-F4, F4-C4, C4-P4, P4-O2), the right temporal chain (FP2-F8, F8-T8, T8-P8, P8-O2), and the midline chain (FZ-CZ, CZ-PZ). Figure 2 depicts this montage on the scalp. A data-provenance detail was resolved during quality control: CHB-MIT EDF files list the bipolar derivation `T8-P8` twice, which EDF readers de-duplicate as `T8-P8-0`/`T8-P8-1`; the first occurrence was adopted as the canonical channel, and channel-name normalisation stripped the trailing suffix so the derivation was matched consistently across all files. All 48 recordings passed an integrity check confirming a 256 Hz sampling rate and the presence of all 18 target channels.

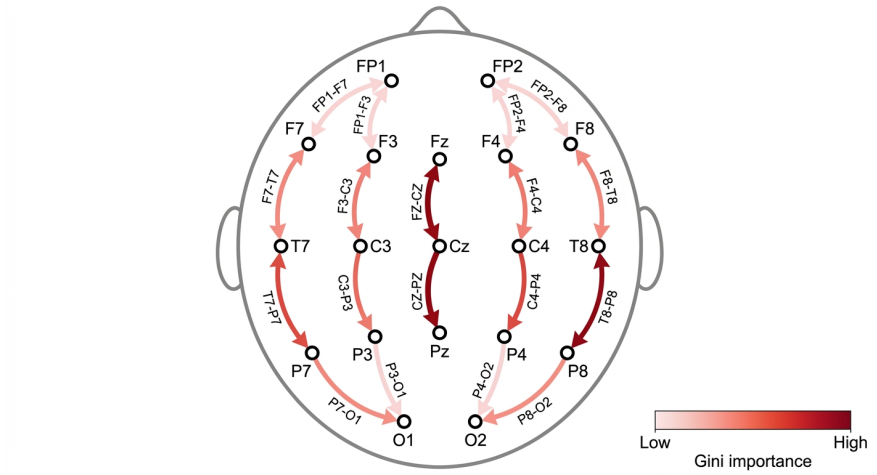

Figure 2. The 18-channel bipolar montage, shaded by learned importance. Each arrow is one bipolar derivation between standard 10–20 electrode positions (nose at top). Shading encodes the cross-fold mean Gini importance summed over the ten features of that channel (darker = more predictive). The central-midline derivations CZ-PZ and FZ-CZ and the right-temporal derivation T8-P8 carry the most seizure-discriminative information, whereas prefrontal derivations contribute least. Quantitative values appear in Figure 7.

2.3 Preprocessing

Each recording was processed with MNE-Python [20]. The 18 target channels were selected and reordered, a 60 Hz notch filter was applied to suppress mains interference, and a 0.5 Hz to 45 Hz finite-impulse-response band-pass filter was applied to retain the physiological frequency range relevant to seizure activity while removing slow drifts and high-frequency noise. Signals were confirmed at 256 Hz (with resampling available had any file differed) and converted to microvolts. This band-pass and notch configuration is consistent with recommended practice for scalp-EEG seizure analysis, where the seizure-relevant spectrum lies below ∼40 Hz [18, 21].

2.4 Windowing and labelling

Continuous signals were segmented into **4 s windows (1024 samples) with 50 % overlap** (a 2 s, 512-sample step). This window length is long enough to estimate spectral band powers at 0.5 Hz resolution yet short enough to localise seizure activity, and the overlap increases the effective sampling of the rare ictal class. Labelling was designed to prevent pre-/post-ictal contamination. In seizure-containing ("ictal") files, only windows lying *entirely* within an annotated seizure interval were retained and labelled ictal (class 1); windows overlapping the uncertain seizure boundaries or the surrounding non-seizure signal were discarded. In seizure-free ("interictal") files, windows spanning the recording were labelled interictal (class 0). To further guard against leakage from the immediate peri-ictal state, ictal and interictal windows were sourced from disjoint files. A post-hoc audit confirmed that all 1372 ictal windows lay strictly inside annotated seizure intervals and that no ictal window originated from a seizure-free file (zero violations). The final dataset comprised 49 959 windows—1372 ictal and 48 587 interictal—an ictal prevalence of 2.75 %, with zero missing or non-finite feature values.

2.5 Feature extraction

From each channel of each window we computed ten features, giving $18 \times 10 = 180$ features per window. Four are time-domain descriptors: *variance* (signal power), *skewness* and *kurtosis*

(distribution shape), and *line length*—the sum of absolute sample-to-sample differences, a computationally cheap measure that jointly captures amplitude and frequency and is a long-established seizure-onset feature [22]. Six are frequency-domain descriptors derived from a Welch power spectral density estimate (512-sample segments, 50 % overlap) [23]: the *relative band power* in the delta (0.5 Hz to 4 Hz), theta (4 Hz to 8 Hz), alpha (8 Hz to 12 Hz), beta (12 Hz to 30 Hz), and gamma (30 Hz to 45 Hz) bands, and the normalised *spectral entropy* over the 0.5 Hz to 45 Hz range, which quantifies how ordered or broadband the spectrum is. This compact, interpretable set spans the feature families most consistently identified as informative for EEG seizure detection [17, 18, 21].

2.6 Classifier

We used a random-forest classifier [24], an ensemble of decision trees that is robust to feature scaling, models non-linear interactions, provides a natural feature-importance measure, and performs strongly on tabular feature-engineered EEG problems while remaining interpretable relative to deep networks [12, 17]. The forest comprised 400 trees with `min_samples_leaf` = 2, `max_features` = $\sqrt{p}$, and balanced class weighting, implemented in scikit-learn [25] with a fixed random seed of 42 for reproducibility.

2.7 Leave-Subjects-Out cross-validation

Because seizures cluster within individuals and recordings of one child are mutually correlated, we evaluated with **six-fold Leave-Subjects-Out Cross-Validation** (Figure 3). In each fold one child was held out entirely as the test set and the remaining five children formed the training set; disjointness of subject identity between training and test was asserted programmatically, so no window from a test child could appear in training. This patient-independent protocol is the appropriate analogue of deploying a detector to a new patient and is markedly more demanding—and more honest—than record-wise or window-wise splitting [17]. To avoid information leakage through preprocessing, feature standardisation (`StandardScaler`) was fit on the training fold only and then applied to both training and test data using the training statistics.

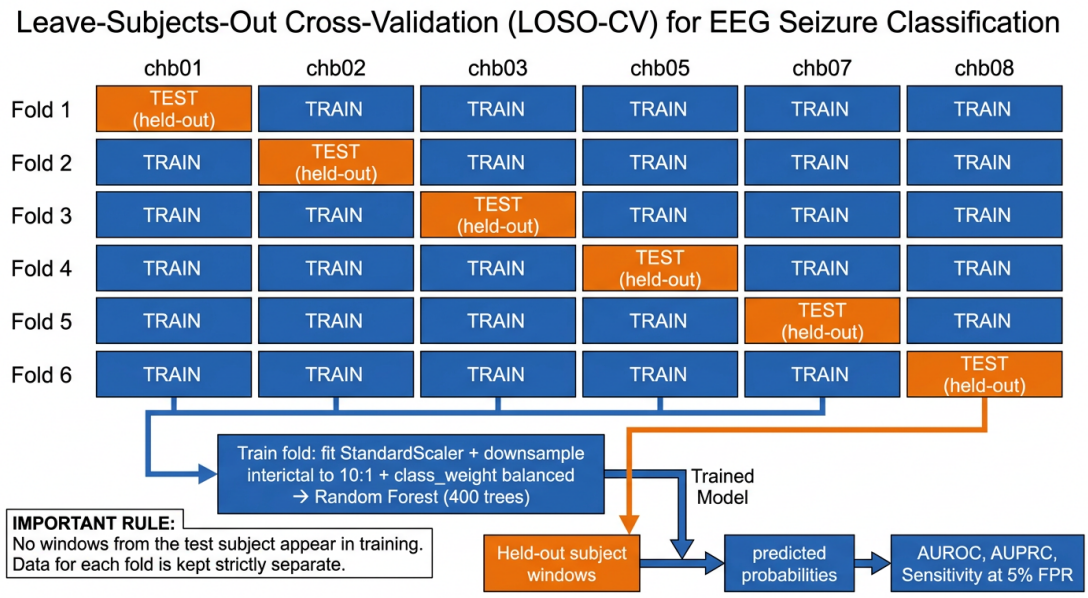

Figure 3. Leave-Subjects-Out cross-validation. In each of the six folds a single child is held out for testing (orange) while the other five provide the training data (blue). Imbalance handling (majority downsampling to 10 : 1 and balanced class weighting) and feature standardisation are applied to the training fold only; the held-out child's windows retain their natural prevalence and are scored to produce per-window probabilities and the reported metrics.

2.8 Class-imbalance handling

With ictal windows forming only 2.75 % of the data, imbalance was addressed explicitly and, crucially, *only on the training folds*. Two complementary techniques were combined. First, the majority (interictal) class was randomly *downsampled to a* 10 : 1 *interictal-to-ictal ratio* within each training fold (seed 42), reducing the dominance of trivially-classified background segments while retaining ten-fold more negatives than positives to preserve a realistic decision surface. Second, *balanced class weighting* was applied inside the classifier so that residual imbalance was compensated in the splitting criterion. The test fold was never resampled or reweighted: it preserved the child's true clinical prevalence, so that all reported metrics reflect performance under realistic operating conditions. This train-only strategy avoids the optimism introduced when synthetic or resampled examples influence the evaluation set [26, 27].

2.9 Evaluation metrics

Under severe imbalance, overall accuracy and even AUROC can present a misleadingly favourable picture, because a detector can score well simply by exploiting the abundant negative class [27, 28]. We therefore report three complementary metrics, both pooled over all held-out windows and per held-out child. *AUROC* (area under the receiver-operating characteristic curve) summarises ranking performance across all thresholds. *AUPRC* (area under the precision–recall curve) is reported against the ictal prevalence baseline, so that the fold in prevalence is made explicit and the lift over chance is interpretable. Finally, because clinical acceptability hinges on false alarms, we fix the operating point at a **5 % false-positive rate** and report the corresponding *sensitivity at* 5 % *FPR*—the fraction of ictal windows recovered when the detector is constrained to raise false alarms on at most 5 % of interictal windows. This fixed-false-alarm framing mirrors how detection devices are judged in practice [9, 11].

3 Results

3.1 Overall (pooled) performance

Pooled across all 49 959 held-out windows from the six LOSO folds, the detector achieved an **AUROC of 0.979** and an **AUPRC of 0.650** (Figure 4). Against the ictal prevalence baseline of 0.0275, the AUPRC represents a 23.7× **lift** over chance, confirming that the ranking quality reflects genuine ictal–interictal discrimination rather than exploitation of the majority class. At the fixed operating point the model recovered **89.4 % of ictal windows at a 5 % false-positive rate** (achieved FPR 0.049 at a probability threshold of 0.294). The precision–recall curve (Figure 4, right) maintains precision above 0.8 up to recall≈0.4 before the expected decline, an operating profile far above the 0.027 prevalence floor. These headline figures are summarised in Table 2.

Table 2. Pooled and macro-averaged LOSO-CV performance. The macro average is the unweighted mean over the six held-out children and is dominated less by the two larger-prevalence subjects than the pooled figure. AUPRC is reported alongside the ictal-prevalence baseline it must be judged against.

Aggregation	AUROC	AUPRC (baseline)	Sensitivity @ 5 % FPR
Pooled (all 49 959 windows)	0.979	0.650 (0.027)	89.4 %
Macro-average (mean over 6 children)	0.987	0.816	93.9 %

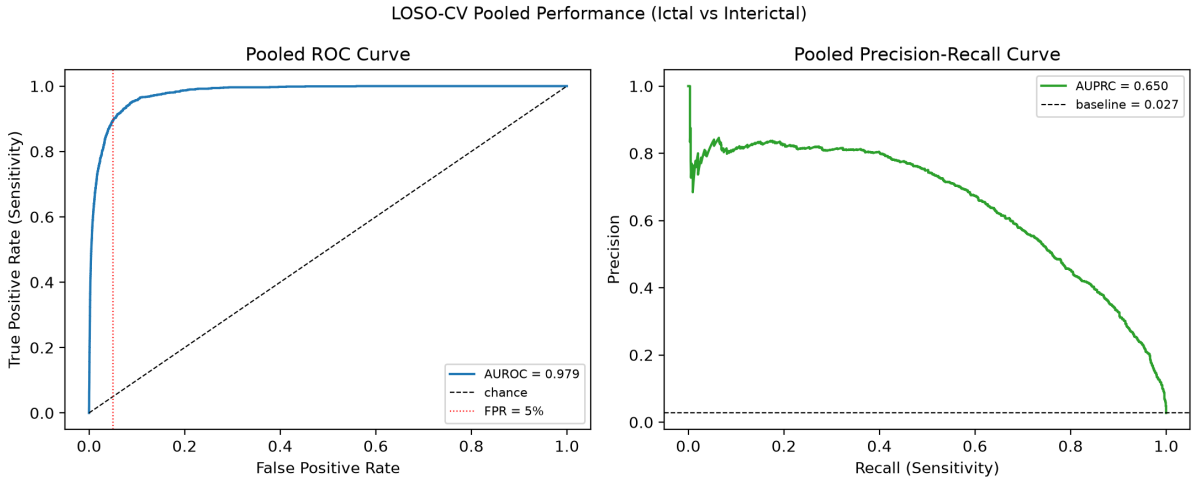

Figure 4. Pooled ROC and precision–recall curves under LOSO-CV. Left: the receiver-operating characteristic (AUROC = 0.979); the dotted vertical line marks the fixed 5 % false-positive operating point at which sensitivity is 89.4 %. Right: the precision–recall curve (AUPRC = 0.650); the dashed line is the 0.027 ictal-prevalence baseline, so the curve sits far above chance across the operating range.

3.2 Per-subject performance and generalisation

Because pooled metrics can be dominated by the subjects contributing the most windows, the per-child results (Table 3, Figure 5) are the more clinically meaningful view: each row is the performance obtained when that child was entirely unseen during training. **Every held-out child exceeded an AUROC of 0.97**, with values ranging from 0.973 (chb03) to 0.998 (chb01). AUROC is thus uniformly high and subject-independent. AUPRC, the more stringent measure under imbalance, showed greater spread: from 0.605 (chb05) to 0.977 (chb01). Notably, this spread does not simply track prevalence—chb07, despite the lowest prevalence (0.73 %), achieved an AUPRC of 0.905 (a 124× lift) and a perfect 100 % sensitivity at 5 % FPR—indicating that

some children's seizures are intrinsically more separable from their background than others'. Sensitivity at the fixed 5 % FPR ranged from 83.9 % (chb03) to 100 % (chb07), with a macro mean of 93.9 %. The consistently high AUROC across all six unseen children is direct evidence that the learned features are not memorising subject-specific signatures but capturing generalisable electrographic correlates of seizure activity.

Table 3. Per-subject performance when each child is the held-out test set. Prevalence is the ictal fraction of that child's windows; the AUPRC lift is AUPRC divided by this prevalence. All metrics are computed on the child's untouched, naturally-imbalanced windows.

Subject	Windows	Prevalence	AUROC	AUPRC	AUPRC lift	Sens. @ 5 % FPR
chb01	5,609	3.78 %	0.998	0.977	25.8×	99.5 %
chb02	5,479	1.50 %	0.988	0.819	54.7×	95.1 %
chb03	5,589	3.44 %	0.973	0.706	20.6×	83.9 %
chb05	5,670	4.81 %	0.980	0.605	12.6×	93.0 %
chb07	21,761	0.73 %	0.998	0.905	123.9×	100.0 %
chb08	5,851	7.76 %	0.984	0.885	11.4×	91.6 %
Mean	—	2.75 %[†]	**0.987**	**0.816**	—	**93.9** %

[†]Pooled prevalence; the per-subject mean prevalence differs.

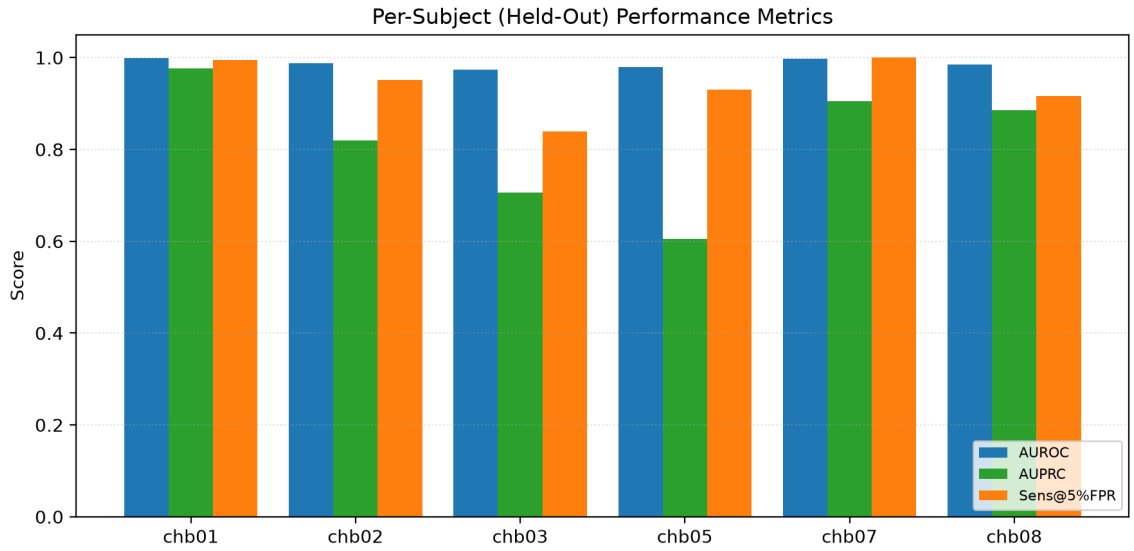

Figure 5. Per-subject held-out performance. AUROC (blue), AUPRC (green), and sensitivity at 5 % FPR (orange) for each child when held out. AUROC is uniformly high (> 0.97); AUPRC varies most, reflecting how separable each child's seizures are from their own interictal background.

4 Feature Importance and Model Interpretation

To understand *what* the model learned—and to check that it relies on physiologically plausible signal properties rather than artefacts—we extracted the random-forest Gini importances from all six LOSO folds and averaged them across folds for a subject-independent view. The per-feature cross-fold standard deviations were small relative to the means, indicating stable rankings that do not depend on which child was held out.

4.1 Which signal properties matter

Aggregating the 180 importances by feature type (Figure 6) reveals a striking concentration in the time domain. **Variance** alone accounted for a summed importance of 0.598, and **line length** for 0.259; together these two amplitude/complexity measures carry roughly **86** % of all predictive importance. This is physiologically coherent: seizures typically manifest as rhythmic, high-amplitude discharges that sharply increase both signal power (variance) and waveform tortuosity (line length), and it echoes the long-standing evidence that line length is an efficient and powerful seizure feature [21, 22]. The spectral features contributed more modestly, led by relative gamma (0.028) and beta (0.026) power—bands associated with fast ictal activity—while spectral entropy and the lower-frequency band powers each contributed around one to two percent. The dominance of simple time-domain measures is an encouraging result for deployment, since variance and line length are trivially cheap to compute in real time on resource-constrained hardware.

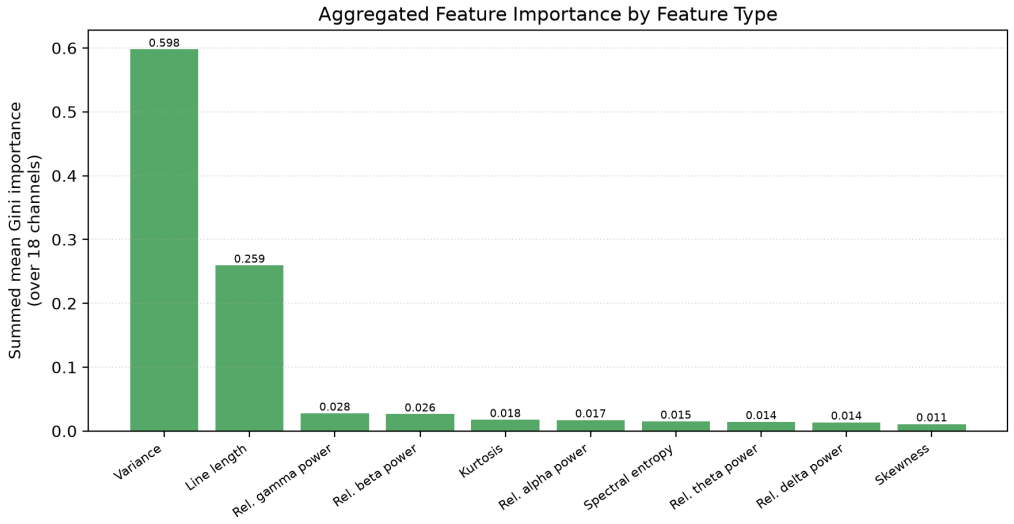

Figure 6. Importance by feature type (summed over 18 channels). Variance and line length dominate, together accounting for ∼86 % of total importance; relative gamma and beta power lead the spectral features.

4.2 Which scalp regions matter

Aggregating importance by channel (Figure 7) shows a clear spatial structure. The two **midline derivations, CZ-PZ (0.100) and FZ-CZ (0.092)**, are the single most informative channels, followed by the right-temporal **T8-P8 (0.086)** and a cluster of central and posterior derivations (C3-P3, F4-C4, P8-O2). Prefrontal derivations (FP1/FP2 chains) were least informative (∼0.029 each), consistent with their susceptibility to ocular and frontalis-muscle artefact and their distance from common temporal and central seizure-onset zones. The prominence of central-midline channels is notable: the vertex region provides a stable, low-artefact vantage on the widespread field changes that accompany many generalised and secondarily-generalised seizures, which aligns with the clinical interest in single- or few-channel detectors placed over central regions [29]. Figure 2 renders these values directly on the scalp.

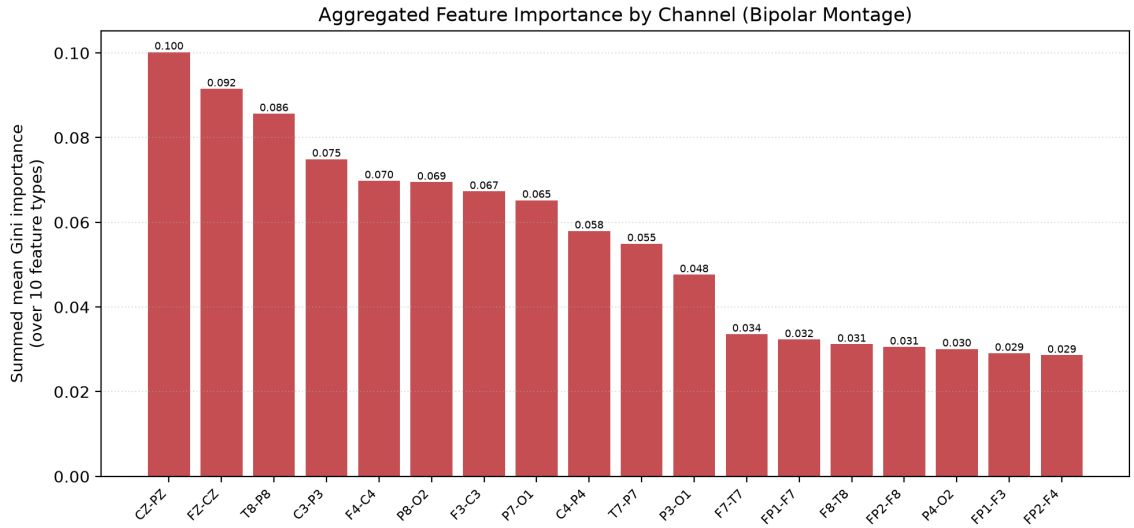

Figure 7. Importance by bipolar channel (summed over 10 feature types). Central-midline (CZ-PZ, FZ-CZ) and right-temporal (T8-P8) derivations are most predictive; prefrontal derivations contribute least.

4.3 Top individual features

The ranking of individual features (Figure 8) unifies the two views above: the ten most important features are almost entirely *variance* and *line length* computed on central and temporal channels. The single strongest feature was `T8-P8__var` (mean importance 0.059), followed by `F4-C4__var`, `CZ-PZ__var`, `C3-P3__var`, and `FZ-CZ__var`. Line-length features on the same midline and posterior channels (`FZ-CZ__linelen`, `CZ-PZ__linelen`, `P7-O1__linelen`) appear immediately below. The cross-fold error bars confirm that these top features remain highly ranked regardless of which child is held out, reinforcing that the detector's decision rule is stable and subject-independent.

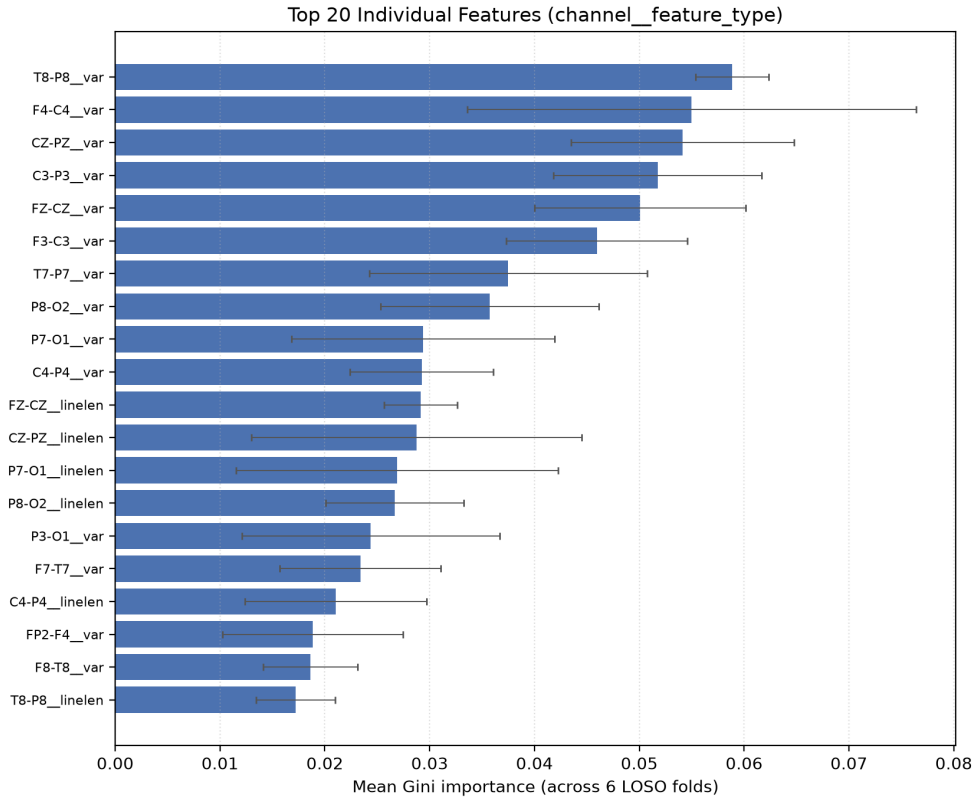

Figure 8. Top 20 individual features (`channel__feature_type`), ranked by mean Gini importance across the six LOSO folds; error bars are the cross-fold standard deviation. Variance and line length on central/temporal derivations dominate, with stable rankings across folds.

5 Discussion

5.1 Principal findings

Under a strict subject-independent evaluation, a compact random forest built on 180 interpretable features detected pediatric scalp-EEG seizures with pooled AUROC 0.979, AUPRC 0.650 (23.7× over prevalence), and 89.4 % sensitivity at a 5 % false-positive rate; every one of the six held-out children exceeded AUROC 0.97. The fact that this performance was obtained *without any data from the test child in training* distinguishes it from the patient-specific paradigm that dominates the CHB-MIT literature and that, by construction, cannot serve a newly-presenting patient [15, 17]. The model's reliance on variance and line length—cheap, causal, time-domain measures—means the detector is both interpretable and computationally light, properties that matter for the wearable and bedside contexts in which automated detection is most needed [9, 22].

5.2 Clinical utility and the false-alarm constraint

Framing performance at a fixed false-alarm rate is deliberate: reviews of seizure detection devices consistently find that clinical and patient acceptance is governed less by sensitivity than by the tolerability of false alarms [10–12]. A window-level sensitivity of 89 %–94 % at 5 % FPR is promising, but the window-level FPR must be translated carefully into an event-level alarm rate: at 256 Hz with 4 s windows stepped every 2 s, a 5 % window FPR corresponds to a substantial number of isolated false-positive windows per hour of monitoring. In practice this is mitigated by post-processing—temporal smoothing, requiring several consecutive positive windows, and

refractory periods—which trades a small amount of latency and sensitivity for a large reduction in false alarms [12, 30]. Because seizures span many consecutive windows whereas artefacts tend to be transient, such smoothing is especially effective; incorporating and tuning it against an event-level false-alarm target is the natural next step toward a deployable detector, and the per-window probabilities released with this report are structured to support exactly that analysis. The premium placed on specificity is exemplified by implanted closed-loop systems such as responsive neurostimulation, whose on-device detectors are deliberately tuned for very low false-detection rates so that therapy is triggered at or just before seizure onset [31].

5.3 Subject-to-subject variability

The most clinically important message of the per-subject analysis is that average performance conceals meaningful heterogeneity. While AUROC was uniformly high, AUPRC ranged from 0.605 (chb05) to 0.977 (chb01)—a spread that reflects genuine differences in how separable each child's seizures are from their own interictal background, driven by seizure type, onset zone, age, and artefact burden. This variability is the central obstacle to a one-size-fits-all detector and explains why patient-specific calibration remains attractive despite its data requirements. A pragmatic middle path, supported by recent patient-independent work, is to deploy a strong subject-independent base model and then lightly personalise it—through threshold re-calibration or brief fine-tuning—once a small amount of labelled data becomes available for a new patient [29, 30]. The uniformly high AUROC we observe suggests that such personalisation would need to adjust operating points rather than re-learn the fundamental seizure signature.

5.4 Relation to prior work

Our design choices align with the methodological direction the field has been urged to take: patient-independent evaluation, explicit imbalance handling, and imbalance-robust metrics [17, 27, 28]. Studies that adopt genuinely subject-independent or leave-one-subject-out protocols on CHB-MIT typically report lower and more variable performance than patient-specific studies, precisely because the task is harder; our results sit at the strong end of that realistic range while remaining transparent about per-subject spread. Compared with recent deep-learning approaches (CNN, LSTM, and hybrid models) that can learn representations directly from raw EEG [32–34], the feature-based random forest sacrifices some potential representational power in exchange for interpretability, low computational cost, stable behaviour on modest data, and a directly auditable decision rule—an appropriate trade-off for a report whose aim is a rigorous, transparent baseline rather than a leaderboard score.

5.5 Limitations

Several limitations bound the interpretation of these results. First, the analysis covers six children from a single centre; although this exceeds the stated six-subject minimum and yields robust LOSO estimates, external validity to other populations, montages, and acquisition hardware is untested, and CHB-MIT's single-centre pediatric nature is a well-known constraint on generalisability [17]. Second, evaluation is at the *window* level; while this is the correct granularity for characterising a classifier, clinicians care about *event*-level detection (did we catch the seizure, and how quickly), which requires the post-processing and onset-latency analysis noted above. Third, the interictal class was drawn from a representative sample of seizure-free files rather than the complete continuous record, so the absolute false-alarm burden over truly continuous multi-day monitoring—including the full diversity of sleep stages, activities, and artefacts—may differ from the controlled window pool used here. Fourth, the discarding of

boundary windows that straddle seizure onset, adopted to prevent label contamination, means the detector is evaluated on established rather than incipient ictal activity and does not directly measure detection latency. Fifth, the handcrafted feature set, though interpretable, may omit discriminative information (for example cross-channel coherence or fine time–frequency structure) that learned representations could exploit. Finally, Gini importance can be biased toward high-cardinality or high-variance features; the strong, stable, and physiologically sensible rankings we observe are reassuring, but permutation importance and ablation studies would provide a more definitive attribution.

5.6 Future work

The immediate extensions follow directly from the limitations: event-level scoring with temporal post-processing tuned to a clinical false-alarm target; onset-latency measurement using boundary-inclusive evaluation; expansion to the full CHB-MIT cohort and to independent datasets (for example adult scalp-EEG corpora) to probe external validity; and controlled comparisons against deep-learning baselines and against lightweight personalisation of the subject-independent model. Each of these builds on the leakage-controlled, imbalance-aware foundation established here.

5.7 Conclusion

A transparent, interpretable, and computationally modest pipeline detects seizures in pediatric scalp EEG with strong performance that *generalises to unseen children*, when—and only when—it is evaluated under a leakage-free Leave-Subjects-Out protocol with explicit imbalance handling and imbalance-robust metrics. Variance and line length over central-midline and temporal derivations carry most of the predictive signal, giving the model a physiologically plausible and auditable basis. The residual per-subject variability, rather than the strong average performance, is the finding most relevant to clinical translation, and it frames the agenda for turning a rigorous baseline into a deployable detector.

Data and Code Availability

The source recordings and clinician seizure annotations are publicly available in the CHB-MIT Scalp EEG Database (version 1.0.0) on PhysioNet [13, 19]. The analysis pipeline—data acquisition, preprocessing and feature extraction, LOSO-CV training and evaluation, and feature-importance analysis—together with the complete per-window prediction table (49 959 rows: subject, file, window index, time span, true label, and predicted ictal probability) accompanies this report. All randomised steps use a fixed seed (42) for exact reproducibility.

References

[1] GBD 2016 Epilepsy Collaborators. Global, regional, and national burden of epilepsy, 1990–2016: A systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, 18(4):357–375, 2019. doi: 10.1016/S1474-4422(18)30454-X.

[2] Kirsten M. Fiest, Khara M. Sauro, Samuel Wiebe, Scott B. Patten, Churl-Su Kwon, Jonathan Dykeman, Tamara Pringsheim, Diane L. Lorenzetti, and Nathalie Jetté. Prevalence and incidence of epilepsy: A systematic review and meta-analysis of international studies. *Neurology*, 88(3):296–303, 2017. doi: 10.1212/WNL.0000000000003509.

[3] Patrick Kwan, Alexis Arzimanoglou, Anne T. Berg, Martin J. Brodie, W. Allen Hauser, Gary Mathern, Solomon L. Moshé, Emilio Perucca, Samuel Wiebe, and Jacqueline French. Definition of drug resistant epilepsy: Consensus proposal by the ad hoc task force of the ILAE commission on therapeutic strategies. *Epilepsia*, 51(6):1069–1077, 2010. doi: 10.1111/j.1528-1167.2009.02397.x.

[4] Patrick Kwan and Martin J. Brodie. Early identification of refractory epilepsy. *New England Journal of Medicine*, 342(5):314–319, 2000. doi: 10.1056/NEJM200002033420503.

[5] Ian O. Miller and Marcio A. Sotero de Menezes. Evaluation and management of drug resistant epilepsy in children. *Seminars in Pediatric Neurology*, 38:100909, 2021. doi: 10.1016/j.spen.2021.100909.

[6] Cynthia Harden, Torbjörn Tomson, David Gloss, Jeffrey Buchhalter, J. Helen Cross, Elizabeth Donner, Jacqueline A. French, Antonio Gil-Nagel, Dale C. Hesdorffer, W. Henry Smithson, Mark C. Spitz, Thaddeus S. Walczak, Josemir W. Sander, and Philippe Ryvlin. Practice guideline summary: Sudden unexpected death in epilepsy incidence rates and risk factors. *Neurology*, 88(17):1674–1680, 2017. doi: 10.1212/WNL.0000000000003685.

[7] Philippe Ryvlin, Lina Nashef, Samden D. Lhatoo, Lisa M. Bateman, Jonathan Bird, Andrew Bleasel, Paul Boon, Arielle Crespel, Barbara A. Dworetzky, Hans Högenhaven, Holger Lerche, Louis Maillard, Michael P. Malter, Cécile Marchal, Jagarlapudi M. K. Murthy, Michael Nitsche, Ekaterina Pataraia, Terje Rabben, Sylvain Rheims, Bernard Sadzot, Andreas Schulze-Bonhage, Masud Seyal, Elson L. So, Mark Spitz, Anna Szucs, Michael Tan, James X. Tao, and Torbjörn Tomson. Incidence and mechanisms of cardiorespiratory arrests in epilepsy monitoring units (MORTEMUS): A retrospective study. *The Lancet Neurology*, 12(10):966–977, 2013. doi: 10.1016/S1474-4422(13)70214-X.

[8] Christian E. Elger and Christian Hoppe. Diagnostic challenges in epilepsy: Seizure under-reporting and seizure detection. *The Lancet Neurology*, 17(3):279–288, 2018. doi: 10.1016/S1474-4422(18)30038-3.

[9] Sándor Beniczky, Samuel Wiebe, Jesper Jeppesen, William O. Tatum, Milan Brazdil, Yuping Wang, Susan T. Herman, and Philippe Ryvlin. Automated seizure detection using wearable devices: A clinical practice guideline of the international league against epilepsy and the international federation of clinical neurophysiology. *Clinical Neurophysiology*, 132 (5):1173–1184, 2021. doi: 10.1016/j.clinph.2020.12.009.

[10] Elisa Bruno, Pedro F. Viana, Michael R. Sperling, and Mark P. Richardson. Seizure detection at home: Do devices on the market match the needs of people living with epilepsy and their caregivers? *Epilepsia*, 61(S1):S11–S24, 2020. doi: 10.1111/epi.16521.

[11] Yaqing Dai, Yiwen Liu, Ke Chen, Lifeng Huang, Ming Song, and Xin Chen. Automated seizure detection with noninvasive wearable devices: A systematic review and meta-analysis. *Epilepsia Open*, 7(2):228–246, 2022. doi: 10.1002/epi4.12588.

[12] Philipp Graetz et al. Minimizing artifact-induced false-alarms for seizure detection in wearable EEG devices with gradient-boosted tree classifiers. *Scientific Reports*, 14:2980, 2024. doi: 10.1038/s41598-024-52551-0.

[13] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.E215.

[14] Ali H. Shoeb. *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2009. Department of Electrical Engineering and Computer Science. Original publication for the CHB-MIT Scalp EEG Database.

[15] Ali H. Shoeb and John V. Guttag. Application of machine learning to epileptic seizure detection. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 975–982, Haifa, Israel, 2010. Omnipress.

[16] Ali Shoeb, Herman Edwards, Jack Connolly, Blaise Bourgeois, S. Ted Treves, and John Guttag. Patient-specific seizure onset detection. *Epilepsy & Behavior*, 5(4):483–498, 2004. doi: 10.1016/j.yebeh.2004.05.005.

[17] Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Xiaodi Huang, and Nasir Hussain. A review of epileptic seizure detection using machine learning classifiers. *Brain Informatics*, 7(1):5, 2020. doi: 10.1186/s40708-020-00105-1.

[18] Turky N. Alotaiby, Saleh A. Alshebeili, Tariq Alshawi, Ishtiaq Ahmad, and Fathi E. Abd El-Samie. A review of feature extraction and performance evaluation in epileptic seizure detection using EEG. *Journal of Ambient Intelligence and Humanized Computing*, 11:1–21, 2020. doi: 10.1007/s12652-019-01679-0.

[19] Ali H. Shoeb. CHB-MIT scalp EEG database (version 1.0.0). PhysioNet, 2010.

[20] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7:267, 2013. doi: 10.3389/fnins.2013.00267.

[21] Lojini Logesparan, Alexander J. Casson, and Esther Rodriguez-Villegas. Optimal features for online seizure detection. *Medical & Biological Engineering & Computing*, 50(7):659–669, 2012. doi: 10.1007/s11517-012-0904-x.

[22] Rosana Esteller, Javier Echauz, Thomas Tcheng, Brian Litt, and Ben Pless. Line length: An efficient feature for seizure onset detection. In *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, volume 2, pages 1707–1710, 2001. doi: 10.1109/IEMBS.2001.1017231.

[23] Peter D. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967. doi: 10.1109/TAU.1967.1161901.

[24] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[26] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

[27] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3): e0118432, 2015. doi: 10.1371/journal.pone.0118432.

[28] Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240, 2006. doi: 10.1145/1143844.1143874.

[29] Taranjit Kaur, Arun Diwakar, Kirandeep, Shivani Chandel, and Tapan K. Gandhi. Single-channel seizure detection with clinical confirmation of seizure onset in long-term scalp EEG. *Frontiers in Neurology*, 15:1389731, 2024. doi: 10.3389/fneur.2024.1389731.

[30] Philippa J. Karoly, Daniel Eden, Ewan S. Nurse, Mark J. Cook, Jodie Taylor, Sonya Dumanis, Caitlin L. Grzeskowiak, and Dean R. Freestone. Removing artefacts and periodically retraining improve performance of neural network-based seizure prediction models. *Scientific Reports*, 13:3466, 2023. doi: 10.1038/s41598-023-30864-w.

[31] Felice T. Sun and Martha J. Morrell. Closed-loop neurostimulation: The clinical experience. *Epilepsia*, 55(11):1779–1786, 2014. doi: 10.1111/epi.12835.

[32] Guangyi Li, Chia-Hung Lee, Jae-Jin Jung, Young Chul Youn, and David Camacho. A one-dimensional CNN-LSTM model for epileptic seizure recognition through EEG signal analysis. *Frontiers in Neuroscience*, 14:578126, 2020. doi: 10.3389/fnins.2020.578126.

[33] Rajesh Singh et al. Epileptic seizure detection from electroencephalogram signals based on 1D CNN-LSTM deep learning model using discrete wavelet transform. *Scientific Reports*, 15:31534, 2025. doi: 10.1038/s41598-025-18479-9.

[34] Jasper H. Bögels et al. Using long short-term memory (LSTM) recurrent neural networks to classify unprocessed EEG for seizure prediction. *Frontiers in Neuroinformatics*, 18:1472747, 2024. doi: 10.3389/fninf.2024.1472747.