Natural Resources for Human Health

eISSN: 2583-1194

Original Research

Received 29 July 2026
Revised 23 Aug 2026
Accepted 07 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (21s): 516–528

KEYWORDS:
ECG, seizure warning, deep learning, candidate heart-rate variability, causal preprocessing, signal availability, developmental evaluation.

ECG-based Seizure Warning: A Developmental Comparison on Public Data

Pravin Andhale[1]*, Varsha H. Patil[2]

[1,2] Matoshri College of Engineering & Research Centre, Eklahare, Nashik, Maharashtra, India

* Corresponding author. E-mail address: pravin.andhale03@gmail.com

ABSTRACT: ECG-based seizure warning requires advance notice, manageable false alarms and sufficient signal availability. We compared a causal waveform convolutional–recurrent network with a heart-rate and candidate-RR logistic baseline on identical SeizeIT2 validation endpoints and alarm rules. A custom partition assigned 50 ECG participants to development, 24 to validation and 50 to reserved testing; one additional participant lacked ECG. The primary protocol used 300-second inputs, 30-second decisions, a five-minute minimum lead, a 20-minute occurrence period and a 25-minute refractory period. Of 72 metadata-eligible validation seizures, 51 had a quality-passing warning opportunity. At threshold 0.5, CNN–GRU warned 22/72 seizures with 1,117 normal-period false alarms over 727.49 scored normal hours; logistic regression warned 21/72 with 1,430 false alarms. Mean patient window AUROC was 0.5273 versus 0.4392 across 16 participants with both classes; the descriptive paired interval included zero. Exploratory short-lead SeizeIT2 models warned 4–8/72 seizures with 4.164–6.825 false alarms per scored normal hour. A separate clean ECG-only Siena experiment reached internal checkpoint selection, but final evaluation and continuous replay were not completed. These developmental results do not establish reliable advance warning, deep-learning superiority or independent cross-dataset generalization. The contribution is a matched-support comparison with explicit alert burden, unavailable opportunities and evaluation provenance.

I. INTRODUCTION

An advance-warning system must do more than classify selected preictal windows. Its input must be available before each decision, its alerts must precede a reference onset by a defined interval, and false alarms must be assessed against explicit exposure. The forecasting literature distinguishes these requirements from retrospective separation of labeled samples [1], [2]. Translation and appropriate reference strategies remain central concerns [3]. This study evaluates ECG waveforms and ECG-derived cardiac features from public recordings; no other biomedical channel is a predictor input.

ECG is motivated by seizure-associated autonomic changes but the presence of a cardiac response does not establish a stable precursor. Reports on autonomic and cardiovascular manifestations and a systematic review of preictal heart-rate changes provide physiological context [4], [5], [6]. Patient-specific HRV and recurrence-quantification analysis offers a related computational approach [7]. Such results cannot be imported as performance expectations when lead times, labeling, processing and exposure differ.

Kalousios et al. reported substantial degradation between retrospective and pseudo-prospective ECG prediction [8]. Weak generalization is therefore not itself a new contribution. Here, the specific contribution is a practical comparison of a causal waveform model and a cardiac-feature baseline on identical validation support, coupled with event availability, complete alarm accounting and explicit selection provenance. The primary question is whether the tested pipelines differ in their observed warning–false-alarm trade-off.

SeizeIT2 provides the primary matched comparison [9]. Model-family, threshold and short-lead analyses are secondary explorations on previously used validation data. Siena contributes a separately specified ECG-only development experiment; it is not a matched replication or a completed external evaluation. A Billeci/Siena provenance audit is supplementary. No reserved-test model outcomes are reported, and the primary validation cohort was used for checkpoint selection and subsequent exploration.

II. RELATED WORK AND STUDY POSITIONING

Seizure-advisory and wearable forecasting studies illustrate the need to evaluate temporal predictions in their intended operating context [10], [11]. These are contextual precedents rather than ECG-only head-to-head baselines. Wearable detection guidance addresses events that have started and is not evidence of pre-onset warning [12]. HRV measurement standards, metric reviews and reporting guidance emphasize requirements beyond a plausible RR series [13], [14], [15]. Candidate RR intervals must be distinguished from independently verified normal-to-normal intervals; spectral balance alone does not identify an autonomic mechanism [16]. Recurrence quantification supplies related background [17], but no recurrence-based predictor is evaluated here. Deep learning in ambulatory ECG and healthcare motivates waveform representation learning without establishing seizure-warning performance [18], [19]. Transfer-learning theory motivates careful consideration of differences between acquisition domains [20], [21]; the present paper does not report a frozen-model cross-dataset transfer experiment. Public-data provenance is supported by PhysioNet, waveform standards, BIDS and OpenNeuro [22], [23], [24], [25], [26]. Siena and the Billeci release require explicit overlap checks rather than being counted as independent cohorts solely because they have different names [27], [28].

III. MATERIALS AND METHODS

A. Cohorts, partitioning and prior exposure

The downloaded SeizeIT2 ECG subset contained 2,804 EDF files from 124 of 125 participants. The custom partition retained the previously exposed 50-participant subset for development. Remaining ECG participants were deterministically assigned to 24 validation and 50 reserved-test participants using a seeded hash ranking stratified by metadata eligibility. All recordings from one participant remained in one role. This is a custom split, not the dataset publication's official split. Table I shows the prequality inventory.

Metadata-eligible events required the declared preceding support and annotation guards; they are a subset of all annotated seizures, not a claim that every source event was usable. The development group included participants previously used as tests in earlier local studies; they are not treated as new test subjects here. Reserved test metadata and sparse engineering information had been analyzed but no reserved test model outcome was used in this paper.

TABLE I. Custom Seizeit2 Partition Before Signal Quality

Role	Patients	EDF files	ECG h	Eligible events
Development	50	568	4,920.55	147
Validation	24	726	1,665.14	72
Reserved test	50	1,510	4,685.24	124
No ECG	1	0	0	0

Counts are based on metadata; all event candidates are subject to signal availability. Results of test models are not reported.

The public-data evidence has distinct roles. SeizeIT2 supplies the primary matched-support comparison. The separate Siena ECG-only arm reached development and internal checkpoint selection, with its evaluation cohort unprocessed in that arm. Known sampled ECG overlap between Siena and the Billeci release prevents treating Billeci as a third independent external cohort. Supplementary Sections S1–S3 retain the protocol and status distinctions. No privately acquired patient recordings are included.

Natural Resources for Human Health
eISSN: 2583-1194

B. Causal waveform preprocessing and availability

Source ECG was processed continuously at 256 Hz with a second-order 0.5-Hz high-pass filter followed by a 129-tap Kaiser-window FIR low-pass filter with a 50-Hz cutoff. Fixed-phase decimation produced 128-Hz waveforms. Filter state persisted across processing chunks and reset at recording boundaries. The 0.25-second FIR delay remained in the signal; predictions were not backdated. At decision time t, the acquired input was the half-open interval [t−300 s, t), following at least 60 seconds of filter warmup. Each input contained 38,400 samples. Observed-window median subtraction and RMS scaling used neither whole-record statistics nor later data. Figure 1 summarizes the shared-support design. Thirty saved engineering checks covered synthetic and two development-recording cases, including chunk equivalence, future-sample perturbation, decimation, acquired-prefix reproduction, finite shape and unavailable-support rejection. These checks support tested implementation properties, not independent beat validation, clinical quality or physiological causation.

C. Shared quality mask and engineered features.

Both primary models used the same experimental endpoint eligibility mask. The screen included candidate intervals of 0.3–2.0 seconds, a past-RR deviation rule, interval-count and coverage constraints, maximum candidate gap, clipping checks and a sustained 200-ms constant-sample check. The five-minute screen required at least 100 retained intervals, at least 80% interval fraction and coverage, and no gap exceeding three seconds. It remains an engineering screen without independently adjudicated beat labels. QRS detection literature provides methodological context [29], [30], while quality-index research and ECGAssess motivate explicit availability accounting [31], [32]. Candidate intervals are not certified sinus NN beats. "Candidate-HRV" therefore denotes the four interval-derived measures. Supplementary Section S5 identifies the saved preprocessing and feature contracts.

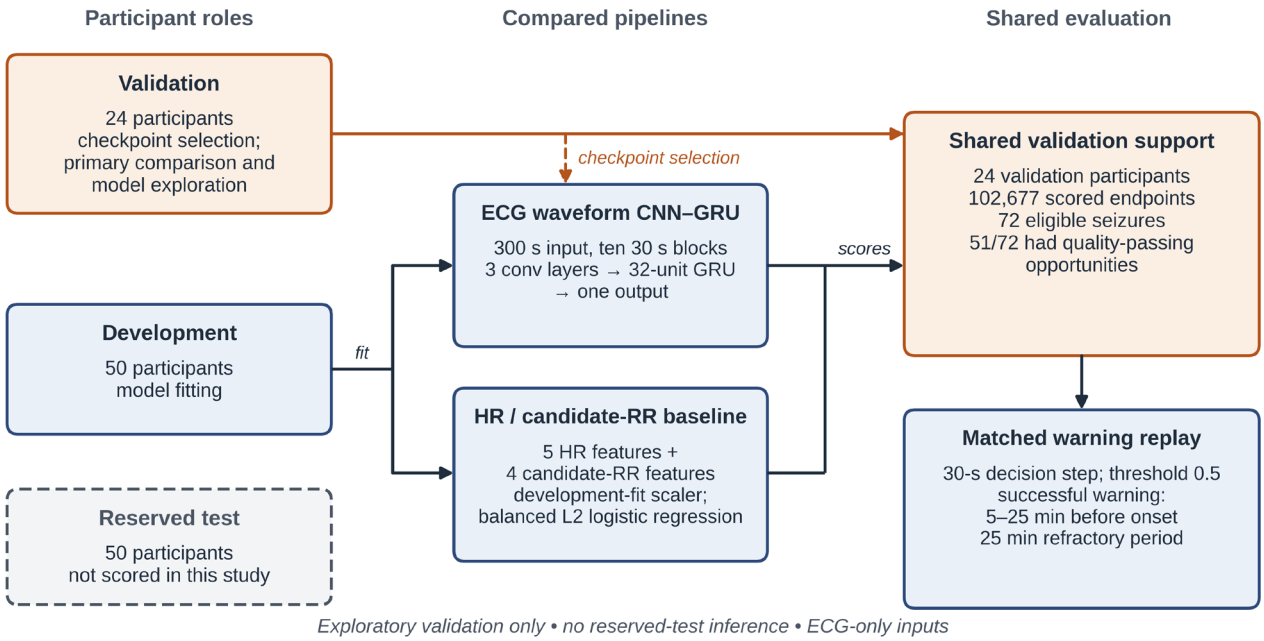


Fig. 1. Study design for the primary ECG-only SeizeIT2 comparison. Both pipelines used identical quality-passing validation endpoints and warning rules. Validation data supported CNN–GRU checkpoint selection and exploratory evaluation; reserved-test model outcomes were not used. One additional participant lacked ECG and was excluded.

The nine baseline features were mean HR, population HR standard deviation, HR slope, minimum HR, maximum HR, mean candidate RR, candidate SDNN, candidate RMSSD and candidate pNN50. The standardization was fitted to the development data only. The balanced L2 logistic regression used C=1 and a maximum of 2,000 optimization iterations. Its training set contained 261,503 labeled development endpoints. It was evaluated at the exact same 102,677 quality-passing validation endpoints as the CNN–GRU, including endpoints with unknown labels during inference.

Natural Resources for Human Health
eISSN: 2583-1194

D. Waveform CNN–GRU and training

The 300-second waveform was arranged as ten 30-second blocks. A shared encoder applied three one-dimensional convolutional layers: 1→16 channels with kernel 7, then 16→32 and 32→64 with kernel 5; all used stride 2 and ReLU activations. Adaptive average pooling and a 64→32 linear projection produced a 32-dimensional embedding per block. A 32-unit GRU summarized the ten embeddings, and a linear head produced one logit. The model contained 21,473 trainable parameters.

The network was trained for 20 epochs using binary cross-entropy with logits, AdamW, learning rate 0.001, weight decay 0.0001 and batch size 16. Each epoch sampled 4,096 development windows with class balance and patient balance within class. The fixed seed was 20261024. No synthetic augmentation was used in the primary experiment. The selected checkpoint was epoch 8, maximizing mean per-patient validation AUROC across participants with both classes, with earliest-epoch tie breaking.

The fitting objectives and sampling differed: logistic regression used all retained labeled development endpoints with class balancing, whereas CNN–GRU sampled classes and participants. This is a practical pipeline comparison, not an isolated estimate of an architecture effect. Sigmoid outputs were not calibrated seizure probabilities.

E. Labels, warning policy and exposure denominators

Supervised labels used the frozen temporal-support rules. Normal supports, including warmup, were excluded if they overlapped annotated onset-minus-one-hour through offset-plus-one-hour guards. Preictal supports excluded other-seizure and artifact contamination according to the saved label contract. Unknown labels were not converted to normal observations. The main replay used a five-minute minimum lead H and a 20-minute occurrence period S. An alert at a matched onset o when $a+H \le o \le a+H+S$, corresponding to 5–25 minutes before onset. Decisions occurred every 30 seconds. After an emitted alert, no further alert could occur for 25 minutes. State reset at recording boundaries; files were not stitched. The primary threshold was 0.5. Table II distinguishes the protocols.

TABLE II. Analysis Protocols

Setting	Input / step	Matching lead	Refractory
Main comparison	300 s / 30 s	5–25 min	25 min
Short-lead source	60 s / 2 s	H to H+60 s	300 s
Siena development only	60 s / 2 s	H to H+60 s; H=10,30,60,120 s	No replay completed

H = 10, 30, 60 or 120 s for the separate short-lead arms. Quality masks, cohorts and completion status differ; the Siena row is not an evaluated warning policy.

Every quality-passing endpoint was scored, including periods excluded from supervised labels. Future annotations did not gate alarm emission; all emitted alerts could affect subsequent refractory timing. Event capture was the number of metadata-eligible seizures with at least one matching alert divided by 72, retaining events without a quality-passing opportunity as misses. Opportunity coverage was reported separately. Normal false-alarm rates used label-zero scored endpoint slots: $N_{normal} \times 30/3600$ hours. These are conditional rates, not false alarms per complete monitoring hour. A separate assessable-horizon analysis required the full future 25-minute horizon, valid annotations and the documented exclusions. Matched, assessable-false and unassessed alerts were retained. The normal-label and assessable-horizon views must not be added together. Neither independently certifies all remaining time as seizure-free. Always-warning and never-warning controls used the same available decisions and refractory rule. Always-warning emitted at the first eligible decision after refractory expiration; it was not an uninterrupted warning state and could capture fewer than all events with an opportunity.

F. Statistical reporting and exploratory analyses

Window AUROC was computed within each validation participant with both classes, then averaged equally across participants. ROC and precision–recall methodology support discrimination assessment [33], [34], but do not alone measure alert utility. The paired AUROC interval resampled 16 participants for 2,000 bootstrap replicates. A manuscript-stage descriptive analysis

resampled all 24 validation participants in pairs for 10,000 replicates (seed 20261027), aggregating their saved event counts, normal alarms and normal scored hours. Percentile intervals describe HRV-minus-CNN differences; no model was refitted or selected during this analysis [35]. These conditional intervals do not correct for validation-based checkpoint selection, repeated exploration or the small cohort [36], [37]. Leakage risks are disclosed rather than assumed absent [38]. No confirmatory p-values or post hoc clinical acceptance thresholds are reported. Additional saved models included random forests, histogram gradient boosting, ECG/HRV fusion and past-baseline-relative features [39], [40]. Five thresholds (0.1, 0.3, 0.5, 0.7, 0.9) were explored for the five-model suite on already inspected validation. Short-lead models used 60-second inputs, four lead targets and a different finite/nonconstant waveform screen; their results do not isolate an effect of reducing lead time. Always/never policies are operational references rather than a full stochastic chance-predictor analysis [41], [42].

G. Separate public-data development and provenance audits

The clean ECG-only Siena arm retained fit participants PN03, PN05, PN07, PN09 and PN10, internal-validation participants PN11 and PN12, and a separately designated, previously exposed retrospective evaluation group PN01, PN06 and PN13. Only ECG samples and ECG-based quality gates were used; an earlier other-channel-intersection experiment is excluded. Its short-lead protocol used 60-second inputs, two-second decisions, minimum leads of 10, 30, 60 and 120 seconds, and a 60-second occurrence period. This differs from the primary SeizeIT2 protocol. No Siena evaluation replay or event-warning performance is reported. The separate Billeci audit used a legacy five-feature classifier with historical future-dependent preprocessing and test-informed checkpoint selection; it is an offline reference, not external validation of the causal CNN–GRU. Its overlap and label-mapping limitations are detailed in Supplementary Section S3.

IV. RESULTS

A. Availability and matched primary comparison

Of the 72 metadata-eligible validation seizures, 51 (70.8%) had at least one quality pass in their warning interval and 21 had none. The main replay scored 102,677 endpoints (or 855.64 hours of 30-second decision slots), with 727.49 of them being normal hours. This slot coverage is about 51.4% of the 1,665.14 recorded validation hours and not 100% continuous. Table III summarizes this event and exposure accounting.

TABLE III. Event and Exposure Accounting, Validation Cohort (Threshold 0.5)

Item	Count / time	Share	Note
Validation participants	24	—	726 EDF files
Recorded ECG	1,665.14 h	100%	all validation files
Metadata-eligible seizures	72	100%	capture denominator
With ≥1 quality-passing decision in warning interval	51	70.8%	opportunity available
Without any quality-passing opportunity	21	29.2%	counted as misses
Warned: CNN–GRU / logistic / always-warning	22 / 21 / 38	30.6% / 29.2% / 52.8%	of 72 events
Scored decision-slot time	855.64 h	51.4%	of recorded hours
Normal-label scored time	727.49 h	85.0%	of scored time; false-alarm denominator
Assessable-horizon time	638.94 h	74.7%	of scored time; separate accounting

Natural Resources for Human Health
eISSN: 2583-1194

Event percentages use the 72 metadata-eligible events; time percentages use the base shown in the Note column. Exposure rows are decision-slot time (102,677 endpoints × 30 s), not all monitoring time.

At threshold 0.5, the CNN GRU warned 22/72 seizures (30.6%) and the logistic baseline warned 21/72 (29.2%). Their normal false alarm rates were 1.535 and 1.966 per scored normal hour, respectively. The always warning control warned 38/72 (52.8%) at 3.320 per scored normal hour; never warning warned none. Table IV shows capture and burden. Figs. 2 and 3 show discrimination and alert outcomes. The primary descriptive comparison did not establish a clinically useful operating point.

TABLE IV. Matched Validation Warning Outcomes At Threshold 0.5

Model	Warned / 72	Normal false	False/h	Median lead, min
CNN–GRU	22/72	1117	1.535	15.01
HR/HRV logistic	21/72	1430	1.966	14.60
Always warning	38/72	2415	3.320	14.90
Never warning	0/72	0	0.000	—

Shared normal scored exposure: 727.49 h. Only 51/72 events had quality-passing opportunities. Median lead is conditional on warned primary events. False-alarm rates are conditional on scored normal endpoint slots (N normal × 30/3600 h), not on complete monitoring time, and must not be multiplied by 24 to get daily rates.

The mean patient AUROC was 0.5273 for CNN-GRU and 0.4392 for logistic regression amongst 16 participants with both usable classes (Fig. 2). HRV-minus-CNN difference was -0.0881 and paired percentile interval [−0.2446, 0.0563]. Patient-specific differences are quite different. This rule does not allow for a claim of superiority from the average alone.

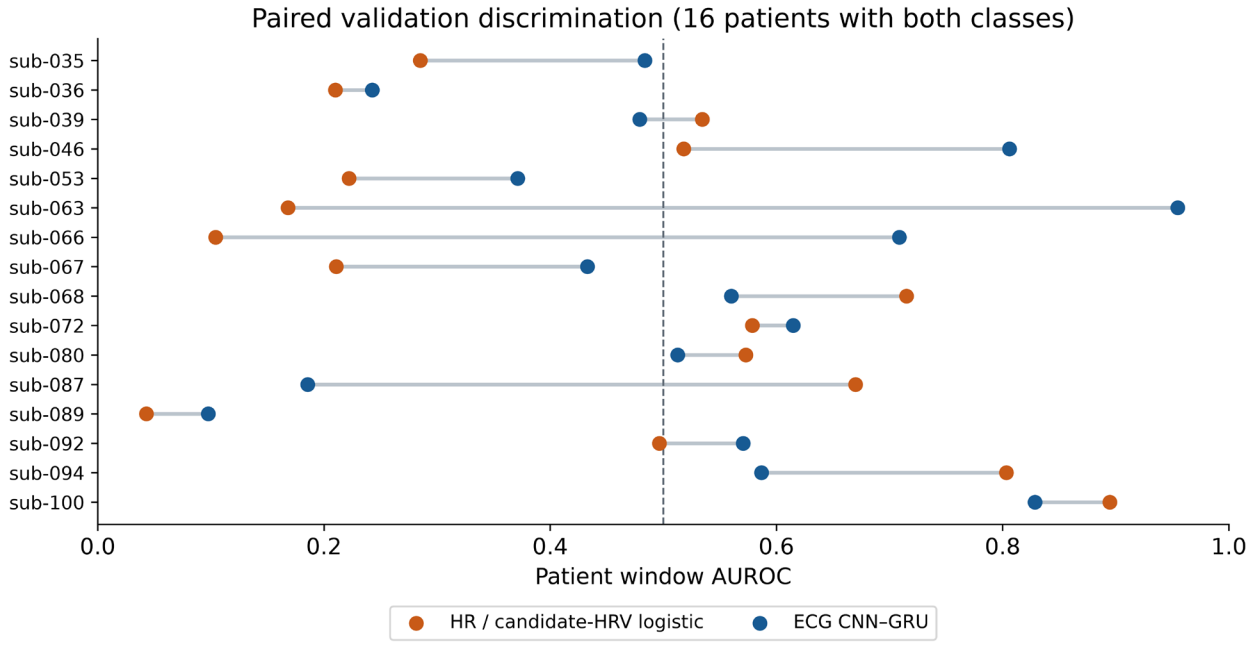


Fig. 2. Paired patient window AUROC for 16 validation participants with both classes. Lines connect results for the same participant. The dashed line is the 0.5 discrimination reference, not a significance threshold. These are checkpoint-selection data.

Natural Resources for Human Health
eISSN: 2583-1194

The post hoc patient-cluster interval for the event-capture difference was also wide, as was the interval for false-alarm-rate difference (Table V). These descriptive intervals condition on selected models and validation exposure. They do not constitute independent confirmation or proof of equivalence.

TABLE V. Paired Descriptive Differences: Hrv Minus Cnn

Outcome	Difference	95% percentile interval
Patient AUROC	−0.0881	[−0.2446, 0.0563]
Event capture, percentage points	-1.39	[-16.84, 16.67]
Normal false alarms/h	0.430	[-0.337, 1.146]

All differences are HRV minus CNN–GRU. AUROC: 16 paired patients, 2,000 replicates. Event/rate: 24 paired patient clusters, 10,000 replicates. Conditional on exposed-validation model selection; not confirmatory intervals.

The CNN–GRU emitted 71 alerts matching valid seizures, 985 assessable-horizon false alerts and 326 unassessed alerts. Logistic regression emitted 64, 1,317 and 330, respectively. These matched-alert counts differ from the primary 22 and 21 warned-seizure counts because alert matching considers valid reference events beyond the metadata-eligible primary subset and may involve multiple alerts. They must not replace the primary event denominator. On 638.94 assessable hours, assessable-false rates were 1.542/h and 2.061/h. These are separate from the guarded-normal rates in Table IV.

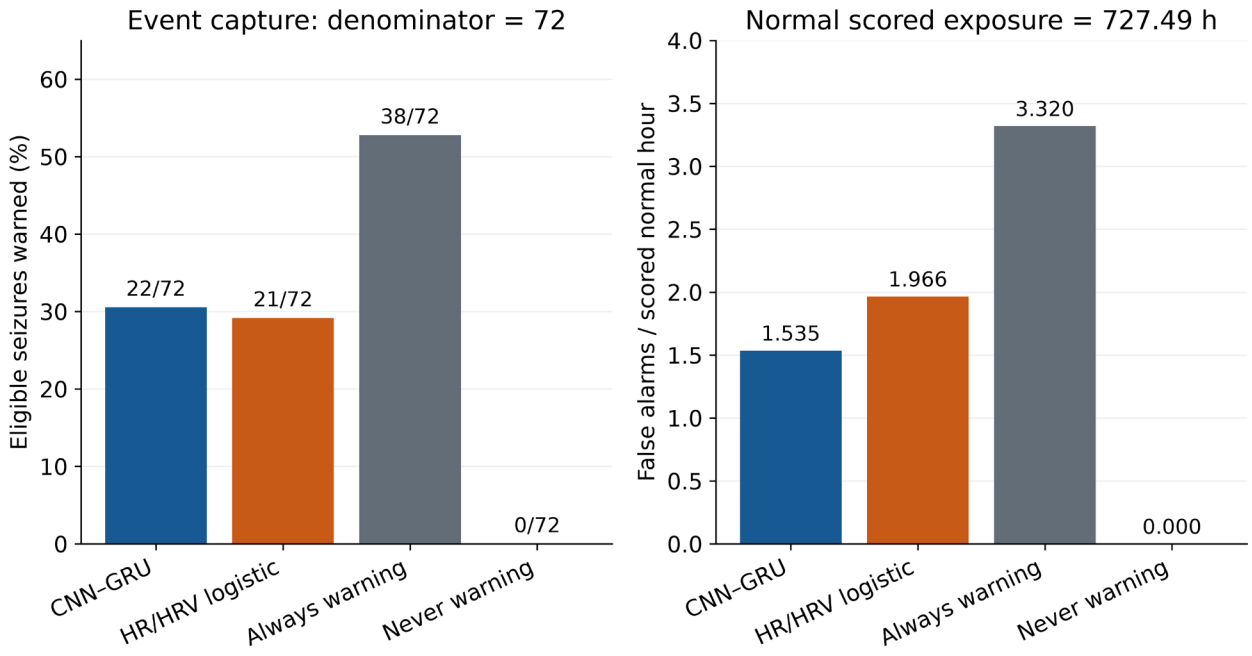


Fig. 3. Fixed-threshold primary validation outcomes. Event capture uses all 72 metadata-eligible events, including 21 without quality-passing opportunities. False-alarm rates use 727.49 scored normal hours rather than all recorded monitoring time.

B. Exploratory alternatives and operating points

Figure 4 retains all five evaluated thresholds for the five-model suite; no operating point is an independent optimum. Mean patient AUROC was 0.4466 for random forest, 0.4283 for histogram boosting and 0.5028 for ECG/HRV fusion. Relative-baseline logistic and boosting models achieved 0.4467 and 0.4652, respectively. At threshold 0.7, HRV logistic warned 6/72 seizures at 0.088 normal false alarms/h, while CNN–GRU warned 10/72 at 0.515/h. A separate relative-baseline search yielded

8/72 at 0.287/h. These estimates arose from validation exploration and do not establish a clinically acceptable operating point. Supplementary Table S1 and the accompanying CSV files retain the public exploratory results.

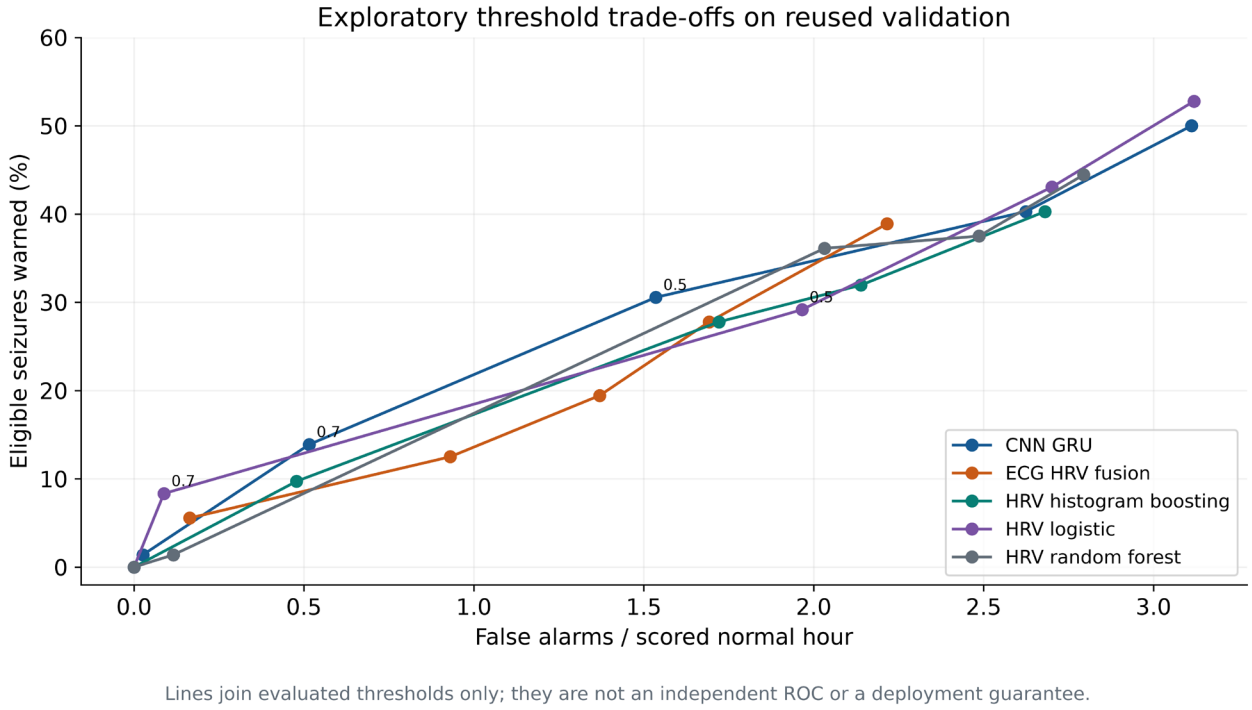


Fig. 4. Five-model/five-threshold operating-point exploration on repeatedly inspected validation data. Annotations identify thresholds for CNN–GRU and logistic regression. Connecting lines are visual guides; thresholds may produce nonmonotonic capture under refractory scheduling.

C. Short-lead exploration

The short-lead SeizeIT2 checkpoint selected epoch 1 with mean patient/lead validation AUROC 0.4661. At threshold 0.5, minimum leads of 10, 30, 60 and 120 seconds warned 6, 8, 4 and 7 of 72 primary events, respectively (Table VI; Fig. 5). Normal false-alarm rates ranged from 4.164 to 6.825 per scored normal hour. At threshold 0.3, capture matched the corresponding always-warning counts; thresholds 0.7 and 0.9 warned zero primary seizures. These outcomes did not establish useful short-lead discrimination.

TABLE VI. Exploratory Short-Lead Validation At Threshold 0.5

Minimum lead	Warned / 72	Normal false	False/h
10 s	6/72	9370	6.825
30 s	8/72	8178	5.957
60 s	4/72	5717	4.164
120 s	7/72	7167	5.220

Normal scored exposure: 1,372.93 h. The screen differs from the main comparison. Occurrence period: 60 s.

Natural Resources for Human Health
eISSN: 2583-1194

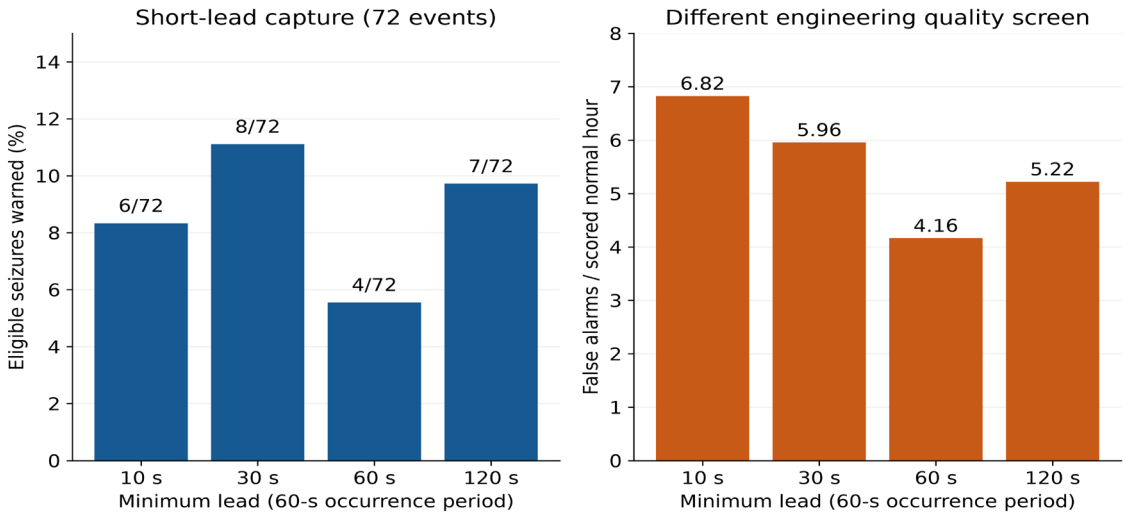


Fig. 5. Short-lead validation at threshold 0.5. Actual warning intervals are minimum lead through minimum lead plus 60 s. The finite/nonconstant waveform screen differs from the main RR-based quality screen, preventing a controlled lead-time comparison.

D. Separate Siena development status and provenance findings

Clean ECG-only Siena preparation covered 17 development recordings with four saved causal/prefix checks. Training completed 20 epochs and selected epoch 14 with mean patient/lead internal-validation AUROC 0.4223 across eight patient–lead pairs. Evaluation waveforms were not processed in that experiment, and continuous replay remains outstanding. This is an internal model-selection result, not a completed independent dataset validation or a directly matched estimate against SeizeIT2. Supplementary Table S2 states the cohort roles, quality rules and completion status.

The Billeci extraction contained 34 EDF files in 15 folders; 29 files in 13 folders contained ECG. Twenty-five files shared sampled ECG prefixes with previously used Siena data. This supports sampled shared signal content, not whole-file identity, and absence of a match does not establish an independent participant. The legacy first-channel reference had both historical label classes for seven participants, with mean patient window AUROC 0.4584. It is neither a replication of the original Billeci predictor nor independent external validation of the present model. Supplementary Section S3 retains this audit as provenance evidence.

Experimental ECG2-minus-ECG1 differencing reduced sampled relative mains-power fractions in three public recordings. Electrode derivation was not independently established, and only one differential-reference participant had both usable classes. Different available supports prevent a representation-level prediction conclusion. These secondary observations are retained in Supplementary Section S4, without inferring warning improvement.

V. DISCUSSION

A. Meaning of the primary comparison

Under shared validation support and a fixed alarm policy, CNN–GRU had numerically greater capture and lower normal false-alarm rate than logistic regression. The difference was one warned seizure, the patient-level intervals included zero, and both models missed most eligible events. These results do not establish deep-learning superiority or clinical utility. Availability accounting is essential: omitting the 21 events without a quality-passing opportunity would conceal a practical limitation. The always-warning control did not capture all events because available decisions and refractory scheduling constrain emission. Neither window discrimination nor high event capture alone demonstrates a useful warning system; false alarms, warning occupancy, timing and coverage must also be assessed.

B. Relationship to previous ECG research

The findings support the need for temporal evaluation highlighted by Kalousios et al. [8], without reproducing that study's cohort or methods. Billeci et al. distinguished initial classification from evaluation of unseen seizures using previous events [7]. Differences in temporal support, feature construction and exposure prevent numerical ranking of heterogeneous reported sensitivities as if they came from one benchmark. The present contribution is the matched-support waveform/feature comparison with explicit availability and alarm accounting. Siena's separate internal-validation result and the Billeci provenance audit do not supply independent confirmation of that primary comparison.

C. Interpretation across public datasets

A second dataset strengthens generalization evidence only when participant independence, compatible inputs and a predefined evaluation are established. Separately trained models under different protocols do not constitute a frozen-model cross-dataset transfer test. Here, SeizeIT2 is the primary developmental comparison; Siena is an explicitly incomplete, separately specified ECG-only development arm. Known Siena/Billeci overlap precludes counting those releases as independent cohorts. No end-to-end clinical deployment, inference-latency, battery-life or daily activity evaluation is inferred from these offline results. The current paper provides no basis for another outcome-driven threshold or architecture search on the same exposed cohorts.

D. Bias, limitations and reporting boundaries

First, checkpoint selection and later exploration reused SeizeIT2 validation outcomes; all headline metrics are developmental. Reserved-test model performance was not evaluated, although prior metadata and engineering exposure are disclosed. Second, candidate RR intervals and engineering quality rules lack independent clinical adjudication; conditional normal-hour rates are not daily false alarms over complete monitoring. Third, patient-cluster intervals condition on selected models and do not remove selection optimism; overlapping windows are not independent events. Fourth, annotation guards do not independently certify seizure absence, and Siena/Billeci overlap and mapping ambiguity limit secondary analyses. Fifth, training sampling differs between the primary pipelines, preventing an isolated architecture-effect claim. Sixth, stochastic null-model testing, probability calibration, full warning-occupancy assessment and prospective clinical evaluation remain incomplete or absent. Reporting is informed by TRIPOD+AI [43]. CONSORT-AI and DECIDE-AI concern later trial or early clinical assessment settings [44], [45]; this retrospective developmental study is not such a trial. Siena final evaluation and replay remain unfinished. Further performance work requires a separately agreed frozen protocol; the reserved test remains closed under the existing stopping rule.

VI. CONCLUSION

We compared a causal ECG CNN–GRU and an HR/candidate-HRV logistic baseline on identical SeizeIT2 validation support and warning rules. They warned 22/72 and 21/72 eligible seizures, respectively, with substantial conditional false-alarm burden and limited signal availability. Exploratory alternatives and shorter lead targets did not establish reliable advance warning. A separate Siena ECG-only experiment reached internal checkpoint selection but not final evaluation. The findings support a transparent developmental comparison; they do not establish deep-learning superiority, independent cross-dataset generalization or the universal impossibility of ECG-based prediction.

DECLARATIONS

Data availability: This study analyzes public SeizeIT2, Siena and Billeci resources cited in the manuscript. Access remains subject to the applicable dataset terms. This manuscript does not redistribute the raw recordings.

Ethics and consent: [AUTHOR VERIFICATION REQUIRED: provide the applicable institutional determination for secondary use of these public de-identified datasets, source-study ethics/consent attribution and any dataset-specific requirements.]

Code and processed results: [AUTHOR VERIFICATION REQUIRED: insert an accessible repository or verified availability statement. No public code release is asserted by this draft.]

Funding, competing interests and author contributions: [AUTHOR VERIFICATION REQUIRED: supply accurate statements approved by all authors.]

Natural Resources for Human Health

eISSN: 2583-1194

Related work and submission: [AUTHOR VERIFICATION REQUIRED: disclose overlap with related manuscripts and confirm author names, order and the submitting/corresponding-author details.]

REFERENCES

[1] F. Mormann, R. G. Andrzejak, C. E. Elger, K. Lehnertz, "Seizure prediction: the long and winding road," Brain, vol. 130, no. 2, pp. 314-333, 2007. doi: 10.1093/brain/awl241

[2] R. G. Andrzejak et al., "Seizure forecasting: Where do we stand?," Epilepsia, vol. 64, no. S3, pp. S62–S71, 2023. doi: 10.1111/epi.17546

[3] P. J. Karoly et al., "Seizure forecasting: The long and winding road to clinical translation," Epilepsia, early view, Jul. 2026. doi: 10.1002/epi.70394

[4] O. Devinsky, "Effects of Seizures on Autonomic and Cardiovascular Function," Epilepsy Currents, vol. 4, no. 2, pp. 43-46, 2004. doi: 10.1111/j.1535-7597.2004.42001.x

[5] L. Mazzola, S. Rheims, "Ictal and Interictal Cardiac Manifestations in Epilepsy. A Review of Their Relation With an Altered Central Control of Autonomic Functions and With the Risk of SUDEP," Frontiers in Neurology, vol. 12, Art. no. 642645, 2021. doi: 10.3389/fneur.2021.642645

[6] E. Bruno, A. Biondi, M. P. Richardson, "Pre-ictal heart rate changes: A systematic review and meta-analysis," Seizure, vol. 55, pp. 48-56, 2018. doi: 10.1016/j.seizure.2018.01.003

[7] L. Billeci, D. Marino, L. Insana, G. Vatti, M. Varanini, "Patient-specific seizure prediction based on heart rate variability and recurrence quantification analysis," PLOS ONE, vol. 13, no. 9, pp. e0204339, 2018. doi: 10.1371/journal.pone.0204339

[8] S. Kalousios et al., "ECG-based epileptic seizure prediction: Challenges of current data-driven models," Epilepsia Open, vol. 10, no. 1, pp. 143-154, 2025. doi: 10.1002/epi4.13073

[9] M. Bhagubai et al., "SeizeIT2: Wearable Dataset Of Patients With Focal Epilepsy," Scientific Data, vol. 12, no. 1, Art. no. 1228, 2025. doi: 10.1038/s41597-025-05580-x

[10] M. J. Cook et al., "Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study," The Lancet Neurology, vol. 12, no. 6, pp. 563-571, 2013. doi: 10.1016/s1474-4422(13)70075-9

[11] M. Nasseri et al., "Ambulatory seizure forecasting with a wrist-worn device using long-short term memory deep learning," Scientific Reports, vol. 11, no. 1, Art. no. 21935, 2021. doi: 10.1038/s41598-021-01449-2

[12] S. Beniczky et al., "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, vol. 132, no. 5, pp. 1173-1184, 2021. doi: 10.1016/j.clinph.2020.12.009

[13] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," Circulation, vol. 93, no. 5, pp. 1043-1065, 1996. doi: 10.1161/01.cir.93.5.1043

[14] F. Shaffer, J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," Frontiers in Public Health, vol. 5, Art. no. 258, 2017. doi: 10.3389/fpubh.2017.00258

[15] S. Laborde, E. Mosley, J. F. Thayer, "Heart Rate Variability and Cardiac Vagal Tone in Psychophysiological Research – Recommendations for Experiment Planning, Data Analysis, and Data Reporting," Frontiers in Psychology, vol. 8, 2017. doi: 10.3389/fpsyg.2017.00213

[16] G. E. Billman, "The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance," Frontiers in Physiology, vol. 4, 2013. doi: 10.3389/fphys.2013.00026

[17] N. Marwan, M. C. Romano, M. Thiel, J. Kurths, "Recurrence plots for the analysis of complex systems," Physics Reports, vol. 438, no. 5-6, pp. 237-329, 2007. doi: 10.1016/j.physrep.2006.11.001

[18] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," Nature Medicine, vol. 25, no. 1, pp. 65-69, 2019. doi: 10.1038/s41591-018-0268-3

[19] R. Miotto, F. Wang, S. Wang, X. Jiang, J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," Briefings in Bioinformatics, vol. 19, no. 6, pp. 1236-1246, 2018. doi: 10.1093/bib/bbx044

[20] S. J. Pan, Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, 2010. doi: 10.1109/tkde.2009.191

[21] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J. W. Vaughan, "A theory of learning from different domains," Machine Learning, vol. 79, no. 1-2, pp. 151-175, 2010. doi: 10.1007/s10994-009-5152-4

[22] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet," Circulation, vol. 101, no. 23, 2000. doi: 10.1161/01.cir.101.23.e215

[23] B. Kemp, A. Värri, A. C. Rosa, K. D. Nielsen, J. Gade, "A simple format for exchange of digitized polygraphic recordings," Electroencephalography and Clinical Neurophysiology, vol. 82, no. 5, pp. 391-393, 1992. doi: 10.1016/0013-4694(92)90009-7

[24] B. Kemp, J. Olivan, "European data format 'plus' (EDF+), an EDF alike standard format for the exchange of physiological data," Clinical Neurophysiology, vol. 114, no. 9, pp. 1755-1761, 2003. doi: 10.1016/s1388-2457(03)00123-8

[25] K. J. Gorgolewski et al., "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments," Scientific Data, vol. 3, no. 1, Art. no. 160044, 2016. doi: 10.1038/sdata.2016.44

[26] C. J. Markiewicz et al., "The OpenNeuro resource for sharing of neuroscience data," eLife, vol. 10, Art. no. e71774, 2021. doi: 10.7554/elife.71774

[27] P. Detti, "Siena Scalp EEG Database (version 1.0.0)," PhysioNet, 2020. doi: 10.13026/5d4a-j060

[28] L. Billeci, D. Marino, L. Insana, G. Vatti, M. Varanini, "Data from: Billeci et al. 'Patient-specific seizure prediction based on heart rate variability and recurrence quantification analysis' (v2)," Zenodo, 2018. doi: 10.5281/zenodo.1415495

[29] J. Pan, W. J. Tompkins, "A Real-Time QRS Detection Algorithm," IEEE Transactions on Biomedical Engineering, vol. BME-32, no. 3, pp. 230-236, 1985. doi: 10.1109/tbme.1985.325532

[30] P. S. Hamilton, W. J. Tompkins, "Quantitative Investigation of QRS Detection Rules Using the MIT/BIH Arrhythmia Database," IEEE Transactions on Biomedical Engineering, vol. BME-33, no. 12, pp. 1157-1165, 1986. doi: 10.1109/tbme.1986.325695

[31] G. D. Clifford, J. Behar, Q. Li, I. Rezek, "Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms," Physiological Measurement, vol. 33, no. 9, pp. 1419-1433, 2012. doi: 10.1088/0967-3334/33/9/1419

[32] L. Kramer, C. Menon, M. Elgendi, "ECGAssess: A Python-Based Toolbox to Assess ECG Lead Signal Quality," Frontiers in Digital Health, vol. 4, Art. no. 847555, 2022. doi: 10.3389/fdgth.2022.847555

[33] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006. doi: 10.1016/j.patrec.2005.10.010

[34] T. Saito, M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," PLOS ONE, vol. 10, no. 3, pp. e0118432, 2015. doi: 10.1371/journal.pone.0118432

[35] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," The Annals of Statistics, vol. 7, no. 1, pp. 1-26, 1979. doi: 10.1214/aos/1176344552

[36] S. Varma, R. Simon, "Bias in error estimation when using cross-validation for model selection," BMC Bioinformatics, vol. 7, no. 1, Art. no. 91, 2006. doi: 10.1186/1471-2105-7-91

[37] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," NeuroImage, vol. 180, pp. 68-77, 2018. doi: 10.1016/j.neuroimage.2017.06.061

[38] S. Kapoor, A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, pp. 100804, 2023. doi: 10.1016/j.patter.2023.100804

[39] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. doi: 10.1023/a:1010933404324

[40] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001. doi: 10.1214/aos/1013203451

[41] D. E. Snyder, J. Echauz, D. B. Grimes, B. Litt, "The statistics of a practical seizure warning system," Journal of Neural Engineering, vol. 5, no. 4, pp. 392-401, 2008. doi: 10.1088/1741-2560/5/4/004

[42] T. Kreuz et al., "Measure profile surrogates: A method to validate the performance of epileptic seizure prediction algorithms," Physical Review E, vol. 69, no. 6, Art. no. 061915, 2004. doi: 10.1103/physreve.69.061915

[43] G. S. Collins et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, pp. e078378, 2024. doi: 10.1136/bmj-2023-078378

[44] X. Liu et al., "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," Nature Medicine, vol. 26, no. 9, pp. 1364-1374, 2020. doi: 10.1038/s41591-020-1034-x

[45] B. Vasey et al., "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," Nature Medicine, vol. 28, no. 5, pp. 924-933, 2022. doi: 10.1038/s41591-022-01772-9