

LookAroundNet: Extending Temporal Context with Transformers for Clinically Viable EEG Seizure Detection

Pór Sverrisson¹ and Steinn Guðmundsson¹

¹University of Iceland, Faculty of Computer Science

January 12, 2026

Abstract

Automated seizure detection from electroencephalography (EEG) remains difficult due to the large variability of seizure dynamics across patients, recording conditions, and clinical settings. We introduce LookAroundNet, a transformer-based seizure detector that uses a wider temporal window of EEG data to model seizure activity. The seizure detector incorporates EEG signals before and after the segment of interest, reflecting how clinicians use surrounding context when interpreting EEG recordings. We evaluate the proposed method on multiple EEG datasets spanning diverse clinical environments, patient populations, and recording modalities, including routine clinical EEG and long-term ambulatory recordings, in order to study performance across varying data distributions. The evaluation includes publicly available datasets as well as a large proprietary collection of home EEG recordings, providing complementary views of controlled clinical data and unconstrained home-monitoring conditions. Our results show that LookAroundNet achieves strong performance across datasets, generalizes well to previously unseen recording conditions, and operates with computational costs compatible with real-world clinical deployment. The results indicate that extended temporal context, increased training data diversity, and model ensembling are key factors for improving performance. This work contributes to moving automatic seizure detection models toward clinically viable solutions.

1 Introduction

Epilepsy is a chronic brain disorder that causes unprovoked seizures due to temporary electrical disturbance in the brain. It affects people of all ages worldwide and arises from various causes, including head trauma, infections, and genetic and environmental factors. However, in many cases, the underlying cause remains unknown, and symptoms often vary considerably between individuals. Epilepsy can greatly affect quality of life, but approximately three-quarters of those diagnosed can manage seizures with medication or surgery [1, 2, 3, 4]. Effective treatment depends on an accurate diagnosis, which is typically based on the detection and analysis of seizures in the electroencephalogram (EEG) signal, a non-invasive recording of the brain's electrical activity captured through scalp electrodes [5].

An EEG recording can vary noticeably depending on the patient, setting and duration of the monitoring [6]. Age plays a significant role, with neonatal, pediatric, and adult EEGs often differing substantially in both appearance and interpretation [7, 8]. Clinical recordings, such as routine EEGs, are performed in a medical setting and monitor the patient for a short period in a controlled environment. In contrast, ambulatory EEGs are often performed at home, providing extended monitoring over several days while the patient continues daily activities, increasing the chances of detecting infrequent epileptic events missed during short clinical visits [9].

Proper diagnosis often requires long-term monitoring, leading to increased use of home-based video-EEG, combining EEG recordings with synchronized video footage of the patient, which reduces the burden on both patients and healthcare systems [10, 11]. However, ambulatory recordings are usually more prone to artifacts, such as muscle movement and electrode displacement, and applying artifact removal techniques in this setting can be challenging [12].

EEGs are typically reviewed through visual examination of a montage, which arranges EEG channels to display brain activity across the scalp. The review process is labor-intensive and time-consuming, requiring expert knowledge and careful analysis to accurately identify seizures from the complex patterns in the EEG signal. Accurate identification often relies on examining the surrounding context, reviewing EEG segments before and after a suspected event to observe gradual changes or sudden shifts that may indicate seizure activity [13, 14]. Variability in seizure appearance across individuals, recording conditions, and equipment further complicates detection [15, 16].

These challenges have spurred growing interest in automated seizure detection methods that use algorithms to efficiently analyze EEG data. Such methods can be classified into real-time and offline approaches, with real-time systems providing immediate alerts for prompt intervention, whereas offline systems can reduce clinicians' workload and expedite the diagnostic process [17]. Over time, research in seizure detection has transitioned from rule-based techniques to the widespread adoption of deep neural networks [18]. These models have demonstrated strong capabilities in analyzing complex patterns, enabling their application across various healthcare

domains [19]. However, despite advancements, automated seizure detection methods still lack the precision and reliability of human specialists [20].

The performance of deep learning models depends heavily on both the amount of training data and the quality of the labels in supervised settings. Therefore, the limited availability of publicly accessible EEG seizure datasets, each with distinct characteristics, poses challenges for developing models that can generalize effectively across different data sources [21]. Curating these datasets is resource-intensive and costly, and their distribution is often constrained by strict regulations surrounding personal health information [22]. Furthermore, existing datasets are typically small in size, often representing recordings from just a handful of patients. This lack of diversity can lead to substantial variability in data distributions, resulting in models trained on a particular dataset facing difficulties when generalizing to more diverse populations [21]. Moreover, disagreements between annotators on the labeling and boundaries of seizure events can introduce inconsistencies and impact model training [23, 24].

Evaluating seizure detection models presents additional challenges due to varying application needs and the lack of standardized validation protocols [25]. In some cases, detecting the exact onset of a seizure is crucial, particularly for real-time interventions. In other cases, identifying the presence of a seizure is more important than detecting its exact location, such as when assessing the effectiveness of pharmacological treatment. These differences lead to varying validation methods and performance metrics across studies, making it difficult to compare results directly. Furthermore, reported performance metrics often do not reflect clinical translatability, as evaluations may omit considerations such as model generalisability, run-time constraints, explainability, and clinically relevant metrics [26].

In this paper, we present a transformer-based model architecture and conduct a comprehensive evaluation across several independent EEG datasets, which vary in clinical settings and patient demographics. The proposed model incorporates surrounding temporal context when classifying EEG segments, and we evaluate the impact of combining training data from multiple sources to enhance generalization. We employ the SzCORE [25] framework for model evaluation and assess the model’s performance on both publicly available and private datasets. In particular, we focus on a large proprietary dataset consisting of ambulatory home recordings to evaluate its clinical relevance.

2 Related Work

Recent research in automatic seizure detection increasingly favors deep learning-based approaches, particularly Convolutional Neural Networks (CNNs) and Transformer architectures. Most CNN-based studies employ a sliding-window strategy, dividing continuous EEG recordings into shorter segments that are processed independently [27, 28, 29]. This al-

lows CNNs to effectively capture spatial patterns and short-term temporal dynamics. For instance, Thuwajit et al. [30] used depthwise convolutions to extract channel-specific features from 4-second EEG segments, followed by spatiotemporal convolutions to model inter-channel dependencies. However, fixed window lengths limit the model’s temporal awareness, making it difficult to capture the progressive development of seizure activity.

In contrast, Chatzichristos et al. [31] introduced a CNN composed of three attention-gated U-Nets, each processing a different view of the EEG signal derived from distinct pre-processing strategies. Their model analyzes entire recordings in a single pass, generating predictions for every time point. Seeuws et al. [32] extended this idea using an event-based training objective that predicts seizure centers and durations. Whole-recording approaches capture long-term temporal dependencies but are computationally intensive for extended EEG monitoring and assume full access to the recording, which limits their suitability for real-time applications.

Effectively capturing temporal context beyond short, localized windows remains a key challenge in EEG analysis. Capturing longer-term dependencies is important to help separate abnormal patterns from background activity, an insight that dates back to the earliest efforts to develop automated seizure detectors [33]. Early approaches, such as Temko et al. [34], proposed averaging the preceding 10 minutes of non-seizure EEG activity to dynamically adjust the seizure probability threshold. Xun et al. [35] proposed a Continuous Bag-of-Words-inspired approach to extract temporal features through context learning. Subsequent work has sought to extend temporal modeling capacity through alternative model architectures [36, 37].

Transformer models [38] have gained traction in EEG analysis due to their self-attention mechanism, which inherently models long-range relationships. Many studies combine convolutional front-ends for local feature extraction with transformer modules to model broader temporal context [39, 40]. Wu et al. [41] achieved good results by extending the U-Net framework with a transformer encoder, using a 60-second window to generate sample-level predictions. Additionally, studies have shown that transformers can scale to very large EEG models with substantial representational capacity [42, 43]. However, this scalability can introduce significant computational demands during both training and deployment.

Although many studies report strong performance, most evaluate their models on a single dataset, which limits the generalizability of their findings. Studies that evaluate on multiple datasets have sometimes demonstrated limited generalization across out-of-sample datasets [25]. These challenges motivate our interest in two core aspects: (1) effectively modeling the temporal context surrounding EEG segments of interest and (2) achieving robust generalization across datasets alongside performance that remains clinically reliable. We believe both aspects are essential for models intended for real-world deployment. Guided by these considerations, this study seeks

Table 1: Datasets used in the study, grouped by training, validation, and test splits.

Dataset	Patients (w/ sz)	# Seizures	Sz Ratio (%)	Duration
<i>Training</i>				
TUSZ	579 (208)	2421	5.34	37d 22h
Kvikna [†]	254 (254)	1099	0.09	589d 07h
<i>Validation</i>				
TUSZ	53 (45)	1081	4.58	18d 03h
Kvikna [†]	36 (36)	243	0.16	66d 07h
<i>Test</i>				
TUSZ	43 (34)	469	5.93	05d 07h
Kvikna [†]	49 (49)	514	0.33	132d 13h
Siena	14 (14)	47	0.56	05d 23h
SeizeIT1	42 (42)	182	0.08	175d 11h

Each dataset lists the number of patients (with seizures), seizures, total recording duration, and the seizure-to-total duration ratio. [†]The Kvikna dataset is proprietary; all other datasets are public.

to address these challenges using transformer-based models.

3 Methods

3.1 Data

The datasets used in this study are summarized in Table 1, with additional details provided in Appendix A. These datasets contain EEG recordings collected from multiple sources and captured in various environments, with recordings ranging from a few minutes to several days of monitoring. All recordings were conducted using the standard 10-20 electrode system and include seizure annotations.

TUSZ

The Temple University Hospital EEG Seizure Corpus (TUSZ) [44] is a dataset created to support research in automatic seizure detection. TUSZ is a subset of a comprehensive collection of clinical EEG recordings from various hospital settings, including ICU and epilepsy monitoring units. Seizures were annotated by trained undergraduates from Temple University through a collaborative review process based on EEG data.

SeizeIT1

SeizeIT1 [45] includes EEG recordings from patients with refractory epilepsy who underwent presurgical evaluation at the University Hospitals Leuven in Belgium. Patients were monitored in a home environment over multiple days and were restricted to move within their room. Seizures were annotated by a neurologist based on video EEG assessments. For 12 seizures, no end time was annotated. For these, we assign a fixed duration of 30 seconds. This matches the approach

used in the SzCORE benchmark and ensures our results are directly comparable.

Siena

The Siena Scalp EEG Database [46] includes EEG recordings collected by the Unit of Neurology and Neurophysiology at the University of Siena, Italy. During the recordings, patients were instructed to remain in bed as much as possible, either asleep or awake. Seizure annotations were performed by expert clinicians using video EEG.

Kvikna

The Kvikna EEG Seizure Corpus v1 is a proprietary subset of ambulatory EEG recordings from Kvikna Medical’s research database. The recordings were collected from patients with suspected epilepsy, who were monitored at home over several days. Patients were permitted to move freely within their homes but were instructed to remain within a camera’s view for most of the time. Two cameras were installed in each home, one in the bedroom and one in the main daytime area. All selected recordings were reviewed and labeled by experts using video EEG.

3.2 Implementation

Model Architecture

Fig. 1 illustrates the architecture of the proposed LookAroundNet transformer. The model takes as input a continuous, multi-channel EEG segment divided into three parts: a central target section, a look-behind section, and a look-ahead section. The look-behind and look-ahead sections together form the context window, which provides temporal context from both past and future signal activity relative to the target segment. This contextual information enables the model to capture temporal dependencies and dynamic patterns that extend beyond the immediate target region. This design mirrors the approach of human reviewers, who interpret EEG signals by integrating observations of preceding and subsequent EEG activity.

The architectural design of LookAroundNet reflects a common pattern in seizure detection models, where temporal dynamics are first extracted from individual EEG channels, followed by spatial modeling to capture inter-channel relationships. The input is first divided into non-overlapping temporal patches, where each patch is a fixed-duration slice from a single channel, mapped via linear projection. We apply a two-layer convolutional feature extractor to the patches. We experimented with both fewer and more layers, but fewer layers noticeably reduced performance, whereas additional layers provided no noticeable improvement.

The patch embeddings are then processed by channel-wise transformer encoder layers, with the temporal order of the patches preserved through learnable positional encodings.

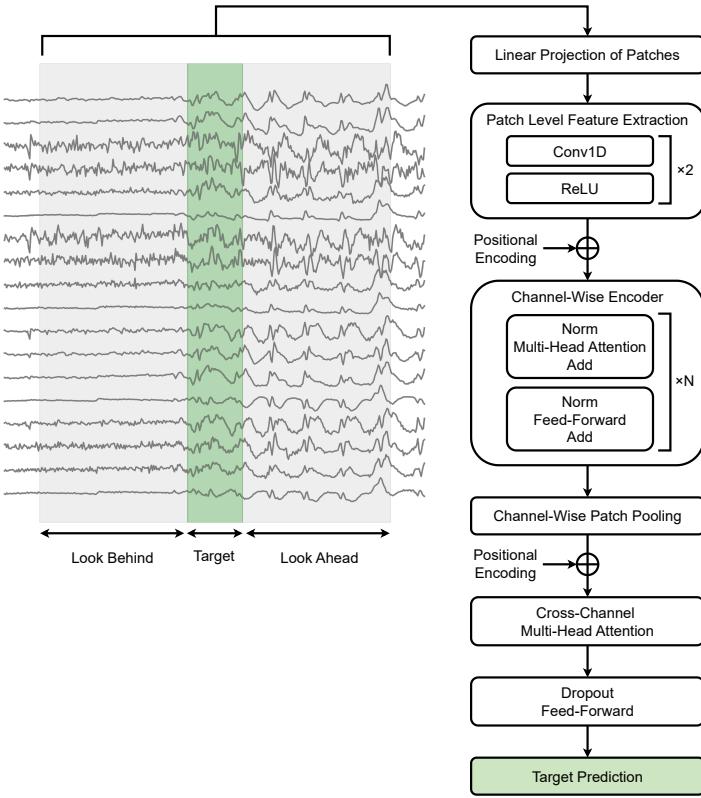


Figure 1: The LookAroundNet architecture. The input window includes a target segment, which is the specific EEG segment to be classified, along with surrounding context segments that provide additional temporal information. Each input channel is independently patched, and features are extracted from each patch. A transformer encoder processes each channel, followed by multi-head attention across channels and a classification layer.

Mean pooling is then applied across the patch dimension, resulting in a single aggregated representation for each channel. A multi-head attention mechanism with learnable positional encodings is applied across channels to capture inter-channel interactions. The resulting representations are passed through a fully connected layer to produce the final class prediction for the target segment.

Data Preparation

The raw EEG signal undergoes pre-processing to standardize its format and improve uniformity across recordings, addressing variations from different settings, sampling rates, and electrode references. First, the signal is transformed into a Longitudinal Bipolar Montage¹. For electrodes included in the montage but missing in the raw recording, their values are estimated by averaging signals from the two spatially closest channels. The signal is then filtered and resampled using the MNE library [47]. This includes FIR filtering with a 0.5 Hz

¹The Longitudinal Bipolar Montage used has the following channels: Fp2-F4, F4-C4, C4-P4, P4-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F7, F7-T3, T3-T5, T5-O1, Fz-Cz, Cz-Pz.

high-pass filter to remove low-frequency drift, a 64 Hz low-pass filter to reduce high-frequency noise, and a 50 or 60 Hz notch filter to eliminate power line interference. Reflective padding is applied at the signal boundaries to minimize edge artifacts from filtering. Finally, the signal is downsampled to 128 Hz.

Experimental Setup

In our experiments, we pre-process the complete datasets before training and evaluation. This setup reflects an offline analysis scenario, as opposed to online processing where smaller data segments would be handled incrementally. In each training epoch, 60k segments are uniformly sampled from the full training set across three categories: fully seizure segments, defined as those with continuous seizure activity across the entire target section; fully non-seizure segments, with no seizure activity in the target section; and mixed segments, where the target section includes both seizure and non-seizure activity. The target segment duration is fixed to 16 seconds for all experiments. To minimize bias toward individual subjects, we sample evenly across patients in the training set, and within each patient, we sample equally from continuous seizure, non-seizure, and mixed portions to balance individual seizure events regardless of their duration.

All models are trained for 200 epochs, and in each epoch, segments are sampled from the full dataset with randomly selected start times. The Kvikna training set comprises very long recordings, spanning almost 600 days. Confidence in the annotations in the portions surrounding marked seizures in this set is high, but there is less confidence in other portions of the recordings. We therefore divide each recording in the Kvikna training set into non-overlapping hour-long segments and retain only those containing at least one seizure for training, discarding the rest. This approach also reduces computational cost and avoids severe class imbalance.

During optimization, segments are assigned a binary label, seizure or non-seizure. For mixed segments, the label is determined by the predominant activity within the target section, defined by whether seizure activity exceeds half of the duration. On the combined TUSZ and Kvikna validation sets, we found that a prediction threshold of 0.5 was suboptimal, as higher thresholds improved the F1-score and reduced false positives, thereby making the model more suitable for clinical purposes. We therefore fixed the threshold at 0.85, and the model achieving the highest validation F1-score at this threshold was selected during training. This model-threshold combination was then used for testing.

All experiments are conducted using four NVIDIA Quadro RTX 5000 GPUs, except for inference time measurements, which use different hardware as described in Section 4.5. The models are trained with CrossEntropyLoss and the AdamW Schedule-Free optimizer [48], using the default parameters. We found that the schedule-free setup stabilizes optimization and prevents divergence during training. We use a cross-entropy loss function with label smoothing, which in our ex-

periments reduced the occurrence of loss spikes during training. We also experimented with gradient clipping, but it resulted in worse performance (data not shown). A complete overview of the model’s architectural and training hyperparameters can be found in Appendix B.

Evaluation

We evaluate the performance of our model across different test sets using the SzCORE framework [25], which is specifically designed to standardize the validation of seizure detection methods. It introduces two primary scoring methods: sample-based and event-based scoring. Sample-based scoring compares predicted and reference labels at a fixed frequency to assess detection precision at the sample level. In contrast, event-based scoring focuses on the overlap between predicted and reference seizure events, providing a more clinically relevant measure. We follow SzCORE recommendations in the choice of evaluation metrics, using sensitivity, F1-score, precision, and false predictions per day (FP/day).

During inference, we iterate over the EEG signal in overlapping segments, advancing the window by 2 seconds at a time. This choice balances computational efficiency with temporal resolution, and predictions from overlapping segments are averaged to generate a continuous binary output. For evaluation, the model output is compared to the ground-truth labels at a frequency of 1 Hz. All reported results are averaged over 5 independently trained models.

4 Results and Discussion

4.1 Model Performance

The results, along with a comparison with selected models from the SzCore benchmark [49], are presented in Table 2. The comparison considers models that performed favorably on the benchmark, including EEG-U-Transformer [41], the winner of the 2025 Seizure Detection Challenge, and EventNet [32]. EventNet is trained on TUSZ, EEG-U-Transformer on TUSZ and Siena, and LookAroundNet uses TUSZ and Kvikna. The LookAroundNet results are reported for two configurations: a single-model setup using a 16-second target window with a 32-second look-behind and a 32-second lookahead, and an ensemble variant described in Section 4.4. SzCORE values for all experiments are provided in Appendix C.

When comparing the single-model variant of LookAroundNet to prior models, the model consistently outperforms existing work on TUSZ and shows overall stronger performance on both Siena and SeizeIT1. Despite having significantly fewer parameters than competing architectures, it achieves higher F1-scores for all test sets, except for the event-based measure on SeizeIT1. Although its performance on this metric is lower than that of EEG-U-Transformer, LookAroundNet compensates with substantially higher event-based sensitivity, detecting many more seizures at the cost of less than one additional false detection per day.

It is important to note that F1-scores do not fully capture the overall performance of a model or its practical applicability. This metric is particularly sensitive to the balance between seizure and non-seizure data, which leads to considerable variation in F1-scores across the datasets in our study. An often more clinically relevant view is to consider the balance between the seizure detection rate (SDR) and the false positive rate. LookAroundNet achieves a strong trade-off between these measures on the Siena and SeizeIT datasets. Performance on Kvikna is somewhat lower, and the false positives per day (FP/day) are comparatively higher for TUSZ.

The reduced performance on the Kvikna dataset likely stems from its composition of long-term home recordings, where the lack of a controlled environment increases the presence of artifacts and signal degradation. Closer inspection revealed that a small number of recordings disproportionately contribute to the false positive rate, primarily due to significant EEG artifacts. These false positives often cluster within segments where signal quality deteriorates, typically caused by loose electrodes or other non-neural disturbances.

Further inspection of the Kvikna dataset revealed that some false positive events were, in fact, genuine seizures that had been missed during the labeling process. No notable differences in performance were observed with respect to the time of day or patient demographics. However, using 16-second segments causes seizures shorter than 8 seconds to go unlabeled, as the classifier requires more than half of a segment to contain seizure activity. In the TUSZ test set, roughly 15% of seizures fall below this threshold. A visualization of the model output for all test sets is provided in Appendix D.

4.2 Impact of Training Set

To study the effect of training data and to compare our results with previous work, we train LookAroundNet on different combinations of public datasets and compare the results to those obtained when the Kvikna dataset is included (see Fig. 2). We examine the seizure detection rate, given by the event-based sensitivity, relative to false positives per day.

For LookAroundNet, we train on TUSZ alone and on TUSZ combined with Siena, SeizeIT1, or Kvikna. Across these experiments, combining TUSZ with Kvikna consistently gave the best results, producing a model that generalized relatively well across all test datasets. In contrast, adding Siena improved the SDR but substantially increased the number of false positives per day. Adding SeizeIT1 slightly improved performance on TUSZ and Kvikna but caused a notable increase in false detections on Siena. The limited benefit observed from incorporating the Siena or SeizeIT1 datasets into the training set likely reflects their relatively small contribution of additional seizures and patients compared with the substantially larger TUSZ and Kvikna datasets.

For a fair comparison to previous studies, we compare LookAroundNet when limited by the same publicly available datasets used in those studies. When trained only on the

Table 2: SzCORE performance metrics for different models across datasets and evaluation methods.

Model	# Params (M)	Metrics	Event based				Sample based			
			TUSZ	Siena	SeizeIT	Kvikna	TUSZ	Siena	SeizeIT	Kvikna
EventNet [32]	1.7	F1-score	59.2	51.7	22.1		44.5	43.7	16.8	
		Sensitivity	53.9	58.7	61.9		44.2	53.2	51.6	
		Precision	43.2	53.7	16.7		33.3	45.5	13.1	
		FP/day	28.4	8.5	19.6					
EEG-U-Transformer [41]	41.0	F1-score	63.6		45.5		40.8		28.2	
		Sensitivity	57.3		37.5		29.3		16.1	
		Precision	49.7		56.2		50.1		63.7	
		FP/day	19.3		0.9					
LookAroundNet (-32, 32) (This study)	0.5	F1-score	72.1	54.9	26.4	21.8	62.1	49.0	31.0	17.6
		Sensitivity	74.2	69.1	58.5	47.4	51.9	42.0	29.3	16.4
		Precision	70.3	46.0	17.1	14.2	77.4	59.4	33.0	18.9
		FP/day	12.7	3.2	1.7	6.5				
LookAroundNet (ensemble) (This study)	1.4	F1-score	77.8	68.5	47.0	31.2	61.8	49.9	37.3	19.9
		Sensitivity	69.2	62.7	49.0	38.4	46.7	37.5	25.4	13.3
		Precision	89.0	75.6	46.2	26.6	91.6	74.5	71.0	40.2
		FP/day	3.4	0.8	0.4	2.5				

The numbers for EventNet and EEG-U-Transformer are obtained from the SzCORE benchmark results [49]. Highest F1-scores for each test set are underlined.

TUSZ dataset, LookAroundNet achieves a significantly higher seizure detection rate and fewer false positives on both TUSZ and Siena, compared to EventNet and EEG-U-Transformer. On SeizeIT1, however, the relative ranking of models is less definitive when considering the trade-off between detection rate and false positives per day. Nonetheless, these results suggest that LookAroundNet’s design and training approach help achieve consistent performance across datasets.

4.3 Context Window Variation

We experimented with shifting the position of the context window relative to the prediction target and varying its duration. Fig. 3 compares model performance when the context window is placed entirely before, centered on, or entirely after the target segment. We report the corresponding event-based F1-scores for each configuration and include a baseline model trained without any surrounding context for reference.

Providing surrounding context consistently improved performance compared to the no-context baseline. This highlights the value of allowing the model to look around the target segment. However, comparing F1-scores across different context window placements shows no consistent advantage for any particular placement. It’s worth noting that incorporating future context introduces a detection delay equal to its duration, since the model must wait for upcoming samples before generating an output. In offline analysis, this delay is not relevant since the full recording is available at any time. For real-time applications, however, it can be preferable to rely on past context to minimize latency.

Fig. 4 shows how model performance varies with the total duration of the context window. Extending the context window generally leads to higher F1-scores, suggesting that incorporating segments surrounding the target segment helps

the model better identify seizure-related patterns. The Siena dataset represents an exception, where increasing the window length shows no additional benefit, and the results exhibit considerable variance due to the limited number of seizures. For the other test sets, the performance gains seem to saturate as the context window lengthens, suggesting that extending the temporal context beyond 128 seconds offers little added value.

4.4 Model Ensembling

Model ensembling can often reduce variance and improve overall performance by combining the predictions of multiple independently trained models. In our experiment, we constructed an ensemble of three LookAroundNet variants, each trained with a different temporal context alignment: 64 seconds before, 64 seconds after, or centered with 32 seconds on either side of the target segment. Predicted probabilities from the models were averaged to obtain the final output. Performance metrics for the ensemble model are provided in Table 2, and F1-score comparisons to the individual configurations are shown in Fig. 3.

Ensembling the configurations resulted in state-of-the-art performance across all datasets, outperforming both the individual models and previously published methods. The ensemble achieved the highest event-based F1-scores and the lowest number of false positives for all test sets. However, the seizure detection rate of the ensemble model was generally lower than that of the individual models. This trade-off, however, can be adjusted by tuning the prediction threshold to favor sensitivity or specificity, depending on the application.

Interestingly, we observed that ensembling different context window configurations yielded larger improvements than simply broadening the temporal context in a single model, as in

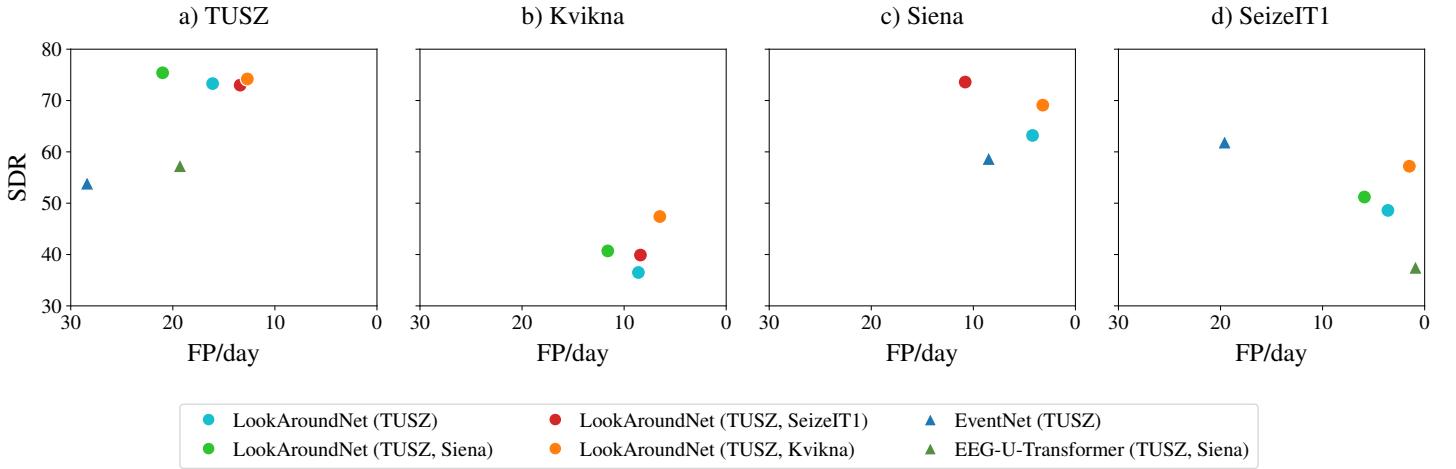


Figure 2: Test set performance of selected models and training set combinations, shown as seizure detection rate versus false positives per day. Each subfigure corresponds to a different test set, with points representing individual model–training set combinations.

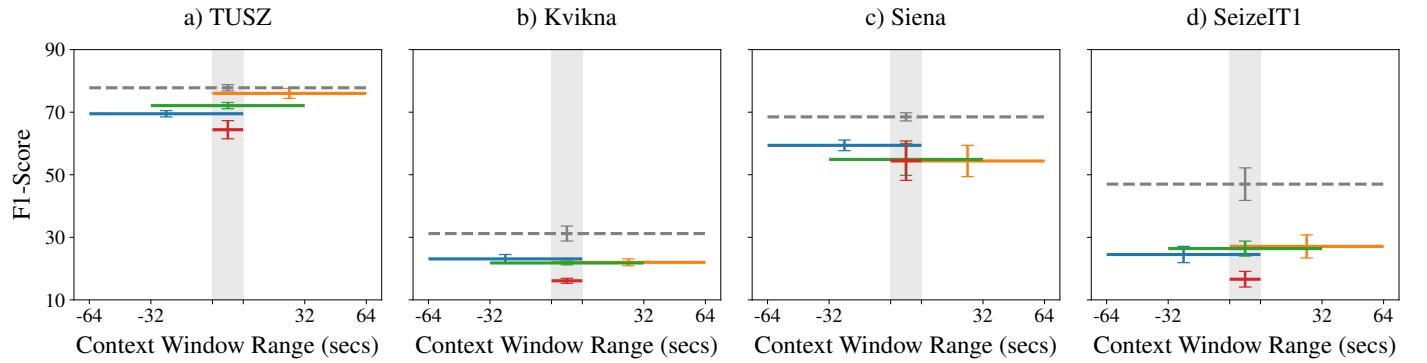


Figure 3: Comparison of event-based F1-scores across test datasets for different context window configurations, each with a total extended context duration of 64 seconds. The highlighted area corresponds to the target segment, and the size and positions of the colored bars represent the context window size and placement. A negative context size indicates past context, whereas a positive size indicates future context. The dashed line represents the ensembled model, created by combining outputs from the three look-around configurations.

Fig. 4. This indicates an advantage of using multiple context window configurations over a single, extended window.

4.5 Cost of Inference

Fast inference is a critical aspect of seizure detection, as it can improve workflow efficiency, enable rapid clinical response, and support real-time monitoring. To evaluate the practicality of LookAroundNet for clinical deployment, we measured its inference performance on one hour of EEG data across several computing platforms. The results of these experiments are summarized in Table 3.

Both an earlier-generation workstation equipped with an Intel Xeon Gold 5218 processor and an NVIDIA Quadro RTX 5000 GPU, and the NVIDIA DGX Spark, a newer desktop-sized AI computer, completed the task in under 6 seconds. In

comparison, a MacBook Pro with an M3 chip took less than 13 seconds. All systems were able to perform inference many times faster than real-time, showing that the model runs efficiently even on standard, general-purpose hardware. Several steps in the pipeline, such as data loading, pre-processing, and inference, can run in parallel to increase throughput. With multiple GPUs, inference can be distributed across devices or used to run multiple models at once, for example, in an ensemble setup.

5 Conclusions

Our evaluation shows that LookAroundNet delivers competitive, state-of-the-art performance on a range of out-of-sample EEG datasets. The model shows strong generalization across diverse clinical settings, demographic groups, and both rou-

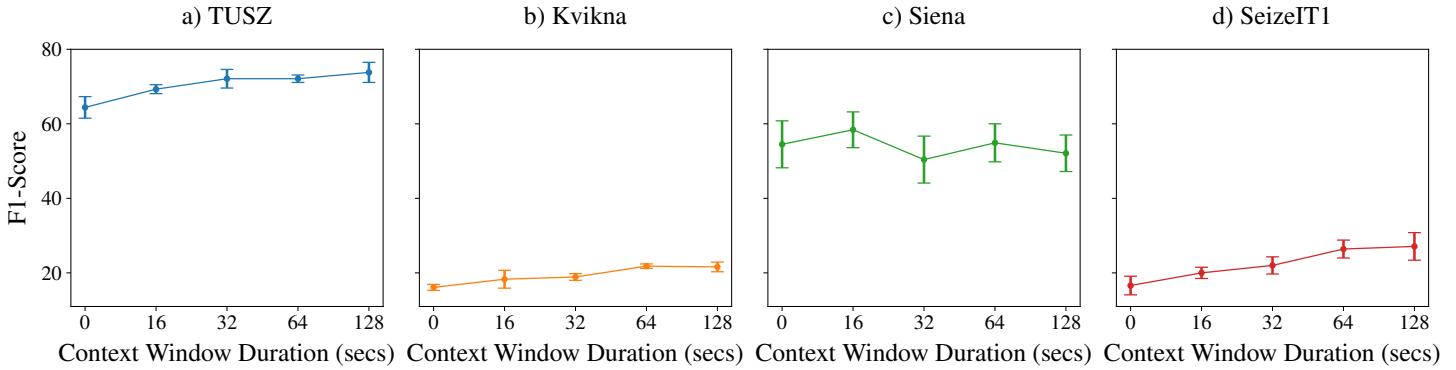


Figure 4: Relationship between event-based F1-score and total context window duration across test sets. Each plot shows results for a model using equal context durations before and after a 16-second target segment.

Table 3: Inference time breakdown across pipeline steps and hardware configurations.

Hardware	Processing Steps (secs)			
	Load Model	Load EEG	Pre-process	Inference
M3 Pro	0.03	0.51	1.09	11.10
DGX Spark	0.05	0.48	1.87	3.09
QRTX 5000	0.12	0.85	1.80	2.64

Results are reported in seconds and correspond to the processing of a 1-hour EDF file using a single model with a total context window size of 64 seconds. Model inference costs about 2 GFLOPs for every second of EEG processed, assuming the inference window advances in 2-second steps.

tine and ambulatory recordings. Furthermore, the results indicate that LookAroundNet has potential for clinical use, offering strong seizure detection performance with relatively low computational requirements.

LookAroundNet demonstrates strong performance compared to prior work when trained solely on the same publicly available datasets used in previous studies. Incorporating data from the proprietary Kvikna dataset of ambulatory recordings further enhances performance, emphasizing the importance of ample and diverse training data in achieving generalization. In addition, extending the model’s temporal context provided an additional boost in detection accuracy, and ensembling models with different context window configurations achieved the best performance.

Although the proposed model demonstrates encouraging performance, several limitations remain. The detection rate on ambulatory data is still relatively low, and certain signal artifacts continue to generate an excessive number of false positives. Moreover, the use of a fixed-length target segment can limit the model’s capacity to capture very short seizure events. These challenges may be alleviated by refining post-processing methods, such as adjusting the prediction threshold, and by enhancing artifact management through targeted training or artifact removal techniques.

Performance varied considerably across datasets, highlighting the impact of differences in recording quality, labeling practices, and clinical context. This reinforces the ongoing challenge of developing seizure datasets with consistent annotations that also capture the diversity of recording environments and patient populations. Consequently, the proposed model may underperform on data types that are underrepresented in this study, such as neonatal recordings.

This study opens up multiple avenues for further research. A worthwhile next step is to better understand the model’s failure modes, including which artifacts cause false predictions and which seizures are often missed. In that context, exploring the model’s internal behavior, such as feature extraction or attention mechanisms, may offer insights into which temporal segments or EEG channels the model relies on for its decisions [50]. This could reveal useful patterns in attention distribution and support improvements in both interpretability and model performance.

For clinical deployment, a lightweight and energy-efficient version of the model is essential to enable real-time seizure detection on edge devices, such as those used in wearable or home-based monitoring systems [51]. Seamless integration with existing EEG platforms and hospital infrastructure is also crucial to support practical workflows, and facilitate clinical analysis. These steps are critical to closing the gap between research and clinical application. As this work suggests, the path to better seizure detection may begin by learning to look around, not only within the data but also toward the clinical environments and technologies that will bring such models to life.

Acknowledgment

We thank the team at Kvikna for their expert assistance in EEG data management and their valuable feedback throughout this project.

Funding

This project received funding from the Icelandic Centre for Research through the Technology Development Fund (grant 2321720-601).

Declaration of Competing Interests

Pór Sverrisson was an employee of Kvikna Consulting ehf. during part of the study period, when Kvikna Consulting ehf. and Kvikna Medical ehf. shared common ownership. Kvikna Medical specializes in the design and development of neurodiagnostic products and solutions under the StratusEEG brand. Steinn Guðmundsson declares no competing interests.

Ethics Statement

The EEG data used in this study had been collected previously with informed consent, anonymised prior to analysis, and processed in accordance with local data-protection regulations. The University of Iceland Science Ethics Committee reviewed the study and found it to be compliant with the scientific ethics rules of the University of Iceland (Ref. SHV2025-113).

Data Availability

The EEG datasets used in this study include both public and proprietary data. Public datasets were obtained from research databases, either freely accessible or subject to data use agreements. Proprietary EEG data were provided by Kvikna Medical ehf. and cannot be made publicly available due to confidentiality agreements.

Code Availability

Code supporting the findings of this study, including scripts for data pre-processing, model training, and evaluation is publicly available at <https://github.com/ths220/lookaroundsnet>.

References

- [1] John S Duncan et al. "Adult epilepsy". In: *The Lancet* 367.9516 (2006), pp. 1087–1100.
- [2] Evangelia Giourou et al. "Introduction to epilepsy and related brain disorders". In: *Cyberphysical Systems for Epilepsy and Related Brain Disorders: Multi-parametric Monitoring and Analysis for Diagnosis and Optimal Disease Management* (2015), pp. 11–38.
- [3] Dieter Schmidt and Steven C Schachter. "Drug treatment of epilepsy in adults". In: *Bmj* 348 (2014).
- [4] Roland D Thijs et al. "Epilepsy in adults". In: *The lancet* 393.10172 (2019), pp. 689–701.
- [5] Shelagh JM Smith. "EEG in the diagnosis, classification, and management of patients with epilepsy". In: *Journal of Neurology, Neurosurgery & Psychiatry* 76.suppl 2 (2005), pp. ii2–ii7.
- [6] W.O. Tatum et al. "Clinical utility of EEG in diagnosing and monitoring epilepsy in adults". In: *Clinical Neurophysiology* 129.5 (2018), pp. 1056–1082. ISSN: 1388-2457. DOI: <https://doi.org/10.1016/j.clinph.2018.01.019>.
- [7] Monika Eisermann et al. "Normal EEG in childhood: from neonates to adolescents". In: *Neurophysiologie Clinique/Clinical Neurophysiology* 43.1 (2013), pp. 35–65.
- [8] Ali Hashemi et al. "Characterizing population EEG dynamics throughout adulthood". In: *ENeuro* 3.6 (2016).
- [9] Jaishree T Narayanan, Douglas R Labar, and Neil Schaul. "Latency to first spike in the EEG of epilepsy patients". In: *Seizure* 17.1 (2008), pp. 34–41.
- [10] William O Tatum, Nimit Desai, and Anteneh Feyissa. "Ambulatory EEG: crossing the divide during a pandemic". In: *Epilepsy & behavior reports* 16 (2021), p. 100500.
- [11] Selim Benbadis et al. "The role of EEG in patients with suspected epilepsy". In: *Epileptic Disorders* 2 (2020), pp. 143–155.
- [12] Jesus Minguillon, M Angel Lopez-Gordo, and Francisco Pelayo. "Trends in EEG-BCI for daily-life: Requirements for artifact removal". In: *Biomedical Signal Processing and Control* 31 (2017), pp. 407–418.
- [13] Michalis Koutroumanidis et al. "The role of EEG in the diagnosis and classification of the epilepsy syndromes: a tool for clinical practice by the ILAE Neurophysiology Task Force (Part 2)". In: *Epileptic Disorders* 19.4 (2017), pp. 385–437.
- [14] Robert S Fisher, Helen E Scharfman, and Marco DeCurtis. "How can we identify ictal and interictal abnormal activity?" In: *Issues in clinical epileptology: A view from the bench* (2014), pp. 3–23.
- [15] Ushtar Amin et al. "Normal variants and artifacts: importance in EEG interpretation". In: *Epileptic disorders* 25.5 (2023), pp. 591–648.
- [16] W David Hairston et al. "Usability of four commercially-oriented EEG systems". In: *Journal of neural engineering* 11.4 (2014), p. 046018.
- [17] Christoph Baumgartner and Johannes P Koren. "Seizure detection using scalp-EEG". In: *Epilepsia* 59 (2018), pp. 14–22.
- [18] Afshin Shoeibi et al. "Epileptic seizures detection using deep learning techniques: a review". In: *International journal of environmental research and public health* 18.11 (2021), p. 5780.
- [19] Andre Esteva et al. "A guide to deep learning in healthcare". In: *Nature medicine* 25.1 (2019), pp. 24–29.
- [20] Elisabeth EM Reus et al. "Automated seizure detection in an EMU setting: Are software packages ready for implementation?" In: *Seizure* 96 (2022), pp. 13–17.

- [21] Sheng Wong et al. “EEG datasets for seizure detection and prediction—A review”. In: *Epilepsia Open* 8.2 (2023), pp. 252–267.
- [22] Tonya White, Elisabet Blok, and Vince D Calhoun. “Data sharing and privacy issues in neuroimaging research: Opportunities, obstacles, challenges, and monsters under the bed”. In: *Human Brain Mapping* 43.1 (2022), pp. 278–291.
- [23] Ana Borovac et al. “Influence of human-expert labels on a neonatal seizure detector based on a convolutional neural network”. In: *Proc. NeurIPS* (2021), pp. 1–5.
- [24] Christa B Swisher et al. “Diagnostic accuracy of electrographic seizure detection by neurophysiologists and non-neurophysiologists in the adult ICU using a panel of quantitative EEG trends”. In: *Journal of clinical neurophysiology* 32.4 (2015), pp. 324–330.
- [25] Jonathan Dan et al. “SzCORE: seizure community open-source research evaluation framework for the validation of electroencephalography-based automated seizure detection algorithms”. In: *Epilepsia* 66 (2025), pp. 14–24.
- [26] Nina Moutonnet et al. “Clinical translation of machine learning algorithms for seizure detection in scalp electroencephalography: systematic review”. In: *arXiv preprint arXiv:2404.15332* (2024).
- [27] Pierre Thodoroff, Joelle Pineau, and Andrew Lim. “Learning robust features using deep learning for automatic seizure detection”. In: *Machine learning for healthcare conference*. PMLR. 2016, pp. 178–190.
- [28] Alison O’Shea et al. “Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture”. In: *Neural Networks* 123 (2020), pp. 12–25.
- [29] Xiaoshuang Wang et al. “One dimensional convolutional neural networks for seizure onset detection using long-term scalp and intracranial EEG”. In: *Neurocomputing* 459 (2021), pp. 212–222.
- [30] PunnaWish Thuwajit et al. “EEGWaveNet: Multiscale CNN-based spatiotemporal feature extraction for EEG seizure detection”. In: *IEEE Transactions on Industrial Informatics* 18.8 (2021), pp. 5547–5557.
- [31] Christos Chatzichristos et al. “Epileptic seizure detection in EEG via fusion of multi-view attention-gated U-Net deep neural networks”. In: *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE. 2020, pp. 1–7.
- [32] Nick Seeuws, Maarten De Vos, and Alexander Bertrand. “Avoiding post-processing with event-based detection in biomedical signals”. In: *IEEE Transactions on Biomedical Engineering* (2024).
- [33] Jean Gotman. “Automatic recognition of epileptic seizures in the EEG”. In: *Electroencephalography and clinical Neurophysiology* 54.5 (1982), pp. 530–540.
- [34] Andriy Temko et al. “Robust neonatal EEG seizure detection through adaptive background modeling”. In: *International journal of neural systems* 23.04 (2013), p. 1350018.
- [35] Guangxu Xun, Xiaowei Jia, and Aidong Zhang. “Detecting epileptic seizures with electroencephalogram via a context-learning model”. In: *BMC medical informatics and decision making* 16.Suppl 2 (2016), p. 70.
- [36] Ramy Hussein et al. “Epileptic seizure detection: A deep learning approach”. In: *arXiv preprint arXiv:1803.09848* (2018).
- [37] Meysam Golmohammadi et al. “Deep architectures for spatio-temporal modeling: Automated seizure detection in scalp EEGs”. In: *2018 17th IEEE international conference on machine learning and applications (ICMLA)*. IEEE. 2018, pp. 745–750.
- [38] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [39] Luca Benfenati et al. “BISeizuRe: BERT-inspired seizure data representation to improve epilepsy monitoring”. In: *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2024, pp. 1–4.
- [40] Yuanda Zhu and May D Wang. “Automated seizure detection using transformer models on multi-channel EEGs”. In: *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2023, pp. 1–6.
- [41] Kerui Wu, Ziyue Zhao, and Bülent Yener. “Large EEG-U-Transformer for Time-Step Level Detection Without Pre-Training”. In: *arXiv preprint arXiv:2504.00336* (2025).
- [42] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. “Large brain model for learning generic representations with tremendous EEG data in BCI”. In: *arXiv preprint arXiv:2405.18765* (2024).
- [43] Jiquan Wang et al. “CBraMod: A criss-cross brain foundation model for EEG decoding”. In: *arXiv preprint arXiv:2412.07236* (2024).
- [44] Vinit Shah et al. “The temple university hospital seizure detection corpus”. In: *Frontiers in neuroinformatics* 12 (2018), p. 83.
- [45] Christos Chatzichristos and Miguel Claro Bhagubai. *SeizeIT1*. Version V1. 2023. DOI: [10.48804/P5Q00J](https://doi.org/10.48804/P5Q00J). URL: <https://doi.org/10.48804/P5Q00J>.
- [46] Paolo Detti. *Siena*. Version v1.0.0. 2020. DOI: [10.13026/5d4a-j060](https://doi.org/10.13026/5d4a-j060). URL: <https://doi.org/10.13026/5d4a-j060>.
- [47] Alexandre Gramfort et al. “MEG and EEG Data Analysis with MNE-Python”. In: *Frontiers in Neuroscience* 7.267 (2013), pp. 1–13. DOI: [10.3389/fnins.2013.00267](https://doi.org/10.3389/fnins.2013.00267).
- [48] Aaron Defazio et al. “The road less scheduled”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 9974–10007.
- [49] Jonathan Dan. *SzCORE - Epilepsy benchmarks*. Accessed: 2025-06-12. 2025. URL: <https://epilepsybenchmarks.com/benchmark/>.

- [50] Dmitry Yu Isaev et al. “Attention-based network for weak labels in neonatal seizure detection”. In: *Proceedings of machine learning research* 126 (2020), p. 479.
- [51] Thorir Mar Ingolfsson. “Robust and Practical Machine Learning Solutions for Wearable Biomedical Edge Devices”. PhD thesis. ETH Zurich, 2025.

Table 4: Overview of the model parameters of the proposed LookAroundNet.

Parameter	Value
Input Patch Size	48
Embedding Dim	96
Convolutional Layers	2
Convolutional Filter (Kernel, Stride)	3, 1
Transformer Encoder Layers	3
Transformer Encoder Feedforward	384
Transformer Encoder Dropout	0.1
Attention Heads	3
Classifier Dropout	0.5

Table 5: Overview of the training parameters.

Parameter	Value
Optimizer	AdamW Schedule-Free
Learning Rate	5e-4
Batch Size	512
Label Smoothing Factor	0.1

A Dataset Details

Fig. 5 reports the age distribution of patients in the training, validation, and test sets. Fig. 6 shows the distribution of seizure durations across the same splits.

B Design Parameters

Table 4 lists the hyperparameters of LookAroundNet. Table 5 summarizes the parameters used during training

C Summary of Results

Table 6 reports SZCORE results for LookAroundNet across different context window configurations, while Table 7 reports results for different training sets.

D Visualization of Model Outputs

Fig. 7 illustrates examples of ensemble model predictions compared to ground truth seizure annotations for test set recordings.

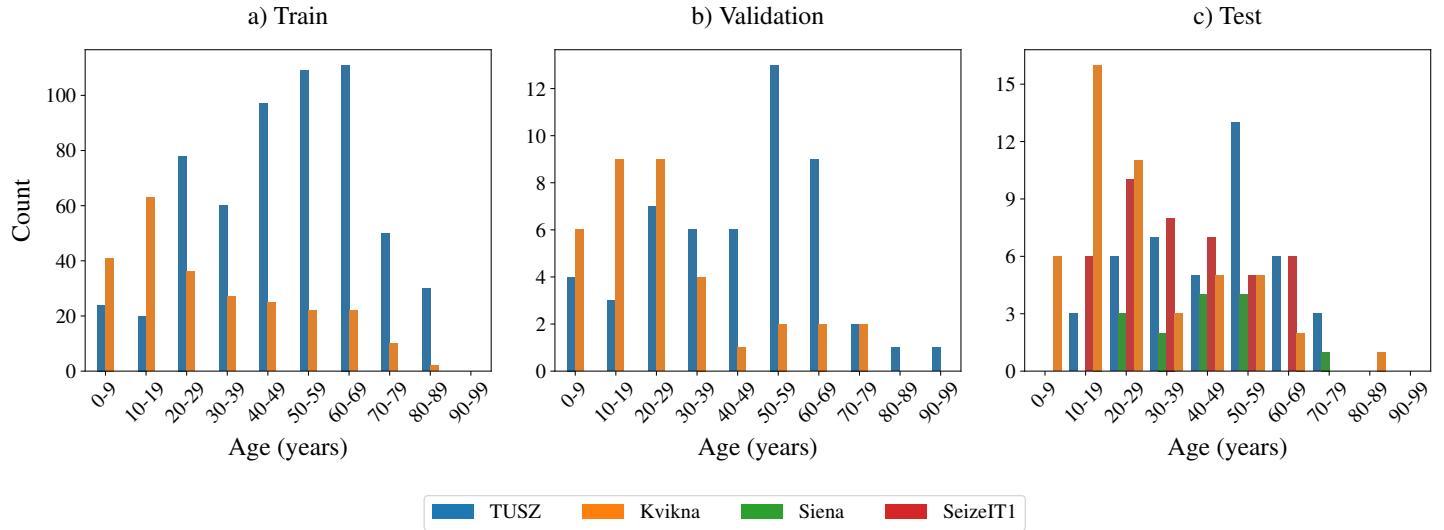


Figure 5: Age distribution of patients in the training, validation, and test sets.

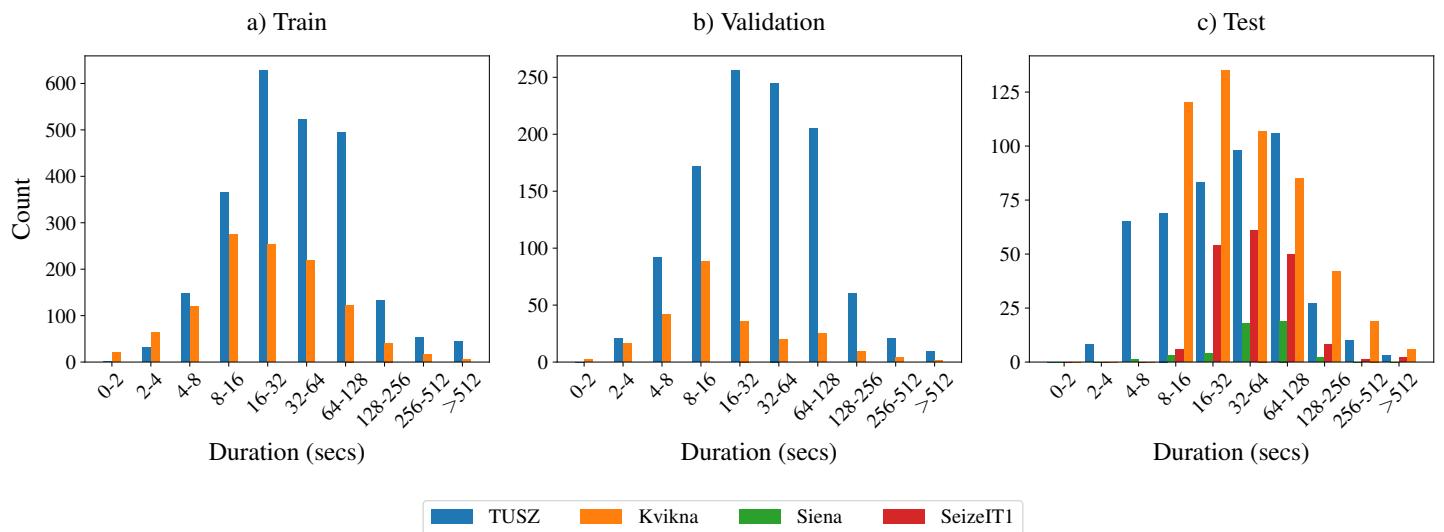


Figure 6: Distribution of seizures by duration across the training, validation, and test sets.

Table 6: Performance of LookAroundNet under SzCORE evaluation across context window configurations.

Context Window Range	Metrics	Event based				Sample based			
		TUSZ	Siena	SeizeIT	Kvikna	TUSZ	Siena	SeizeIT	Kvikna
-0, 0	F1-score	64.4(29)	54.5(63)	16.6(25)	16.1(8)	45.9(39)	42.6(33)	22.3(21)	12.2(6)
	Sensitivity	58.2(29)	55.9(57)	37.1(33)	35.2(24)	33.0(41)	32.0(33)	20.3(10)	12.4(7)
	Precision	72.3(39)	54.8(114)	10.8(21)	10.4(6)	76.5(31)	64.4(67)	25.1(49)	12.1(8)
	FP/day	9.0(15)	1.9(11)	1.8(5)	6.8(5)				
-8, 8	F1-score	69.3(12)	58.4(48)	20.0(15)	18.3(24)	53.4(11)	48.6(24)	23.0(18)	13.3(19)
	Sensitivity	67.9(16)	59.5(41)	47.1(21)	42.2(15)	41.4(18)	37.5(25)	23.6(3)	13.4(9)
	Precision	70.8(22)	58.0(89)	12.7(12)	11.7(20)	75.8(39)	69.2(45)	22.7(39)	13.7(38)
	FP/day	11.2(13)	1.7(7)	1.9(2)	7.4(16)				
-16, 16	F1-score	72.1(25)	50.4(63)	22.0(23)	18.9(9)	54.0(58)	48.3(23)	26.8(16)	14.9(8)
	Sensitivity	71.0(55)	65.0(30)	52.5(28)	44.2(27)	41.4(64)	40.6(23)	25.5(7)	14.9(13)
	Precision	73.5(18)	41.3(73)	13.9(18)	12.0(8)	78.8(46)	59.6(43)	28.5(36)	14.9(14)
	FP/day	10.3(16)	3.6(8)	1.9(3)	7.4(9)				
-32, 32 ⁽¹⁾	F1-score	72.1(10)	54.9(51)	26.4(24)	21.8(6)	62.1(28)	49.0(16)	31.0(18)	17.6(6)
	Sensitivity	74.2(19)	69.1(34)	58.5(22)	47.4(30)	51.9(26)	42.0(14)	29.3(12)	16.4(9)
	Precision	70.3(30)	46.0(72)	17.1(21)	14.2(6)	77.4(31)	59.4(65)	33.0(35)	18.9(10)
	FP/day	12.7(21)	3.2(10)	1.7(3)	6.5(7)				
-64, 64	F1-score	73.8(27)	52.1(49)	27.1(37)	21.6(13)	66.2(36)	48.0(10)	32.2(33)	18.6(19)
	Sensitivity	73.2(15)	75.9(12)	58.5(48)	50.9(27)	54.9(34)	42.7(20)	29.5(24)	19.1(16)
	Precision	74.4(44)	39.8(55)	17.7(30)	13.7(10)	83.4(49)	55.2(37)	36.4(72)	18.2(23)
	FP/day	10.2(24)	4.4(10)	1.6(4)	7.3(6)				
-64, 0 ⁽²⁾	F1-score	69.5(10)	59.4(17)	24.5(26)	23.1(14)	59.6(46)	49.0(23)	29.0(18)	18.0(20)
	Sensitivity	70.1(20)	71.8(26)	55.0(14)	48.4(10)	48.5(49)	39.8(25)	26.4(5)	16.9(20)
	Precision	68.8(14)	50.8(25)	15.8(22)	15.2(13)	77.6(44)	63.8(32)	32.5(53)	19.8(38)
	FP/day	12.8(10)	2.6(3)	1.7(3)	6.2(7)				
-0, 64 ⁽³⁾	F1-score	76.0(16)	54.4(50)	27.1(37)	22.0(11)	64.9(16)	49.8(15)	31.1(12)	18.2(5)
	Sensitivity	72.8(13)	65.5(37)	51.9(8)	44.5(18)	52.4(20)	41.4(15)	27.3(14)	16.9(14)
	Precision	79.6(20)	46.7(64)	18.5(34)	14.7(11)	85.4(39)	62.5(34)	36.3(36)	19.9(18)
	FP/day	7.5(8)	2.9(7)	1.4(3)	5.9(7)				
Ensemble (1), (2), (3)	F1-score	77.8(10)	68.5(13)	47.0(52)	31.2(24)	61.8(34)	49.9(20)	37.3(11)	19.9(6)
	Sensitivity	69.2(13)	62.7(34)	49.0(13)	38.4(11)	46.7(36)	37.5(20)	25.4(9)	13.3(8)
	Precision	89.0(15)	75.6(23)	46.2(104)	26.6(38)	91.6(35)	74.5(8)	71.0(82)	40.2(60)
	FP/day	3.4(5)	0.8(1)	0.4(2)	2.5(5)				

TUSZ and Kvikna datasets are utilized for training.

Table 7: Performance of LookAroundNet under SzCORE evaluation for various training sets.

Training set	Metrics	Event based				Sample based			
		TUSZ	Siena	SeizeIT	Kvikna	TUSZ	Siena	SeizeIT	Kvikna
TUSZ	F1-score	68.7(11)	46.1(41)	12.6(10)	14.2(12)	60.2(35)	45.2(36)	18.3(18)	11.5(12)
	Sensitivity	73.3(21)	63.2(63)	48.6(31)	36.5(14)	49.6(37)	37.9(37)	23.8(18)	12.2(6)
	Precision	64.6(12)	36.6(47)	7.2(6)	8.8(9)	76.8(43)	56.3(47)	15.2(29)	11.0(16)
	FP/day	16.1(10)	4.2(10)	3.6(1)	8.6(8)				
TUSZ, Siena	F1-score	66.2(8)		8.8(10)	12.6(25)	61.9(18)		14.6(16)	11.9(20)
	Sensitivity	75.4(11)		51.2(20)	40.7(29)	52.7(44)		27.2(28)	14.5(13)
	Precision	59.0(16)		4.8(6)	7.5(17)	75.5(42)		10.1(17)	10.1(22)
	FP/day	21.0(16)		5.9(5)	11.6(17)				
TUSZ, SeizeIT1	F1-score	70.7(30)	33.3(72)		16.1(26)	60.4(40)	35.9(64)		12.7(15)
	Sensitivity	73.0(13)	73.6(55)		39.9(26)	49.5(41)	44.3(38)		13.5(12)
	Precision	68.7(47)	22.0(65)		10.2(22)	77.8(49)	32.2(121)		12.5(34)
	FP/day	13.4(26)	10.8(45)		8.4(27)				

LookAroundNet uses a context window configuration of (-32, 32).



Figure 7: Examples of model predictions versus actual seizures for test set recordings using the ensemble model. Correctly predicted seizures are shown in green, with the predicted segment highlighted. Missed seizures are shown in orange, and false predictions in black. Each line represents one or more recordings from a single patient. The recordings may be discontinuous and are stretched to fit the plot width. As a result, equal-length segments on the plot may correspond to different actual durations across patients.