SeizureTransformer on TUSZ: A 27-137x Performance Gap Between Claims and Reproducible Evaluation

true

September 2025

Abstract

SeizureTransformer reports ~1 false alarm per 24 hours on the EpilepsyBench Dianalund dataset. Despite being trained on the Temple University Hospital Seizure (TUSZ) dataset, it has not been evaluated on TUSZ using Temple's official scoring software. We provide, to our knowledge, the first such evaluation with NEDC v6.0.0 and find a 27-137x gap between benchmark claims and clinical reality.

We evaluate the authors' pretrained model on TUSZ v2.0.3's held-out set (865 files, 127.7 hours) and assess identical predictions with four scoring methodologies. With NEDC OVERLAP, the model produces 26.89 FA/24h; with SzCORE, 8.59 FA/24h (~3.1x lower due solely to scoring tolerances); with NEDC TAES, 136.73 FA/24h.

When tuned toward deployment goals, the model cannot meet clinical thresholds with NEDC scoring: targeting 10 FA/24h achieves only 33.90% sensitivity, far below the 75% sensitivity goal for clinical systems (Roy et al., 2021). Acceptable false-alarm rates occur only under SzCORE's permissive tolerances.

We contribute a reproducible NEDC evaluation pipeline, operating points tailored to clinical targets, and quantitative evidence that scoring choice alone drives multi-fold differences. Dataset-matched, clinician-aligned evaluation is essential for credible seizure-detection claims.

1 Introduction

Epilepsy affects 50 million people worldwide, with approximately 30% experiencing drug-resistant seizures that require continuous EEG monitoring for optimal management. The promise of automated seizure detection has tantalized the field for decades, with recent deep learning advances suggesting we may finally be approaching clinical viability. SeizureTransformer, winner of the 2025 EpilepsyBench Challenge, achieved remarkable performance with just 1 false alarm per 24 hours on the Dianalund dataset—a result that would meet stringent clinical deployment standards. Yet when we evaluated this same model on the dataset it was trained on, using the official scoring software designed for that dataset, we discovered false alarm rates 27 to 137 times higher than claimed, depending on the scoring methodology employed.

SeizureTransformer represents a significant architectural advance in seizure detection, combining U-Net's proven biomedical segmentation capabilities with Transformer self-attention mechanisms to capture both local and global EEG patterns. The model was trained on a subset of the Temple University Hospital Seizure (TUSZ) dataset (v1.5.2; ~910 hours) together with the Siena Scalp EEG Database (~128 hours). TUSZ itself is the largest publicly available seizure corpus, but the model's training used a subset as reported by the authors. Its victory in EpilepsyBench 2025, achieving 37% sensitivity at 1 FA/24h on the Dianalund dataset, established it as the state-of-the-art in automated seizure detection. The authors openly shared their pretrained weights, enabling reproducible research and clinical validation.

Despite this success, a critical evaluation gap exists. While SeizureTransformer was trained on TUSZ's train-

ing set, no published evaluation exists using TUSZ's carefully designed, patient-disjoint held-out evaluation set. This 127.7-hour test set, containing 865 files from 43 patients with 469 seizures, was specifically created to enable valid performance assessment. Moreover, Temple University developed the NEDC (Neural Event Detection Competition) scoring software explicitly to match TUSZ's annotation philosophy, ensuring consistent evaluation standards. The absence of TUSZ evaluation is not unique to SeizureTransformer—it reflects a broader pattern where models trained on datasets are evaluated elsewhere, with results reported using varying scoring methodologies.

The choice of scoring methodology profoundly impacts reported performance. The seizure detection community employs multiple evaluation standards, each serving different purposes. NEDC implements strict temporal precision matching Temple's conservative annotation guidelines. In contrast, SzCORE—the Epilepsy-Bench standard—adds 30-second pre-ictal and 60-second post-ictal tolerances around ground truth events, designed to reward clinically useful early warnings. These philosophical differences are not matters of right or wrong but reflect different priorities: research precision versus clinical utility. However, when the same predictions can yield anywhere from 8.59 to 136.73 false alarms per 24 hours purely based on scoring choice, the lack of standardized reporting becomes problematic.

We present, to our knowledge, the first evaluation of SeizureTransformer on TUSZ's held-out test set using Temple's NEDC v6.0.0 scoring software. Our systematic comparison evaluates identical model predictions using four scoring methodologies: NEDC TAES (time-aligned event scoring), NEDC OVERLAP (binary any-overlap), our Python implementation of OVERLAP (achieving perfect parity with NEDC), and SzCORE. At the paper's default parameters (threshold=0.8, kernel=5, duration=2.0s), we observe 45.63% sensitivity at 26.89 FA/24h with NEDC OVERLAP—a 27-fold increase from the Dianalund benchmark claim. The same predictions yield 136.73 FA/24h with NEDC TAES (137-fold increase) and 8.59 FA/24h with SzCORE. This 3.1-fold difference between NEDC OVERLAP and SzCORE stems entirely from scoring methodology, independent of model architecture or parameters.

Our contributions extend beyond revealing performance gaps. We provide: (1) a reproducible NEDC v6.0.0 evaluation pipeline for TUSZ, bridging the research-to-clinic evaluation gap; (2) comprehensive operating points for clinical deployment, including evaluation at a clinically-motivated threshold of $\leq$10 FA/24h; (3) quantitative evidence that scoring methodology alone can account for multi-fold performance differences, highlighting the critical need for transparent reporting; and (4) open-source infrastructure enabling the community to replicate and extend our evaluation framework. When optimizing for the 10 FA/24h threshold, SeizureTransformer achieves only 33.90% sensitivity with NEDC OVERLAP, falling far short of the 75% sensitivity goal for clinical systems (Roy et al., 2021).

The remainder of this paper is organized as follows. Section 2 provides background and related work on TUSZ, NEDC, and scoring methodologies. Section 3 details our evaluation methodology, including data preparation, model inference, and multi-scorer validation. Section 4 presents comprehensive results across multiple operating points and scoring methods. Section 5 discusses implications for clinical deployment, the need for standardized evaluation, and limitations of current benchmarking practices. Section 6 concludes. Section 7 outlines reproducibility resources and exact rerun procedures.

2 Background and Related Work

The Temple University Hospital Seizure Corpus (TUSZ) is the largest publicly available seizure dataset (Shah et al., 2018). Critically, TUSZ implements patient-disjoint train/dev/eval splits—no patient appears in multiple splits—preventing data leakage and enabling valid generalization assessment. The evaluation set contains 865 EDF files totaling 127.7 hours from 43 patients with 469 seizures, specifically reserved for final

held-out testing. This careful split design follows machine learning best practices often violated in medical AI applications. The annotations, performed by board-certified neurologists at Temple University Hospital, follow conservative clinical guidelines requiring clear electrographic seizures with definite evolution and temporal precision in marking onset and offset.

Alongside TUSZ, Temple University developed the Neural Event Detection Competition (NEDC) scoring software suite, creating a matched evaluation ecosystem. NEDC v6.0.0 provides the definitive scoring implementation for TUSZ evaluation. This matched pairing is no coincidence—the same research group created both the dataset and its evaluation tools, ensuring consistency between annotation philosophy and scoring methodology. NEDC implements multiple scoring modes, with OVERLAP (any-overlap binary scoring) serving as the commonly reported evaluation mode for TUSZ. The software is widely used in the literature and serves as a reference implementation for seizure detection evaluation.

The choice of scoring methodology profoundly impacts reported performance, as different methods serve distinct clinical and research priorities. Time-Aligned Event Scoring (TAES), proposed by Shah et al. (2021), represents the strictest evaluation standard, computing partial credit based on temporal overlap percentage—a 60-second seizure with 45 seconds correctly detected receives 0.75 true positive credit. TAES emphasizes temporal precision, making it ideal for algorithm development and research applications where exact timing matters. In contrast, OVERLAP scoring, which NEDC implements as a primary mode, treats any temporal overlap between prediction and ground truth as a full true positive. Shah et al. (2021) note that "OVLP is considered a very permissive way of scoring since any amount of overlap between a reference and hypothesis event constitutes a true positive," yet this binary approach has become a de facto standard for TUSZ reporting, balancing clinical relevance with research needs.

At the most permissive end of the spectrum, SzCORE (Dan et al., 2024) extends any-overlap scoring with clinical tolerances designed for real-world deployment. The system adds 30-second pre-ictal and 60-second post-ictal windows around each ground truth event, recognizing that early warnings before seizure onset provide clinical value and that EEG patterns normalize gradually after seizure termination. Additionally, SzCORE merges predictions separated by less than 90 seconds into single events, substantially reducing alarm fatigue in clinical settings. These modifications, while clinically motivated, can reduce reported false alarm rates by factors of 3-10x compared to stricter scoring methods. Importantly, these different approaches represent not right or wrong methods but rather different valid perspectives on what constitutes meaningful seizure detection—research precision versus clinical utility versus deployment practicality.

SeizureTransformer (Wu et al., 2025) exemplifies both the advances and evaluation gaps in modern seizure detection. The architecture combines U-Net's biomedical segmentation capabilities with Transformer self-attention to capture local and global EEG patterns. Trained on a subset of TUSZ v1.5.2 (~910 hours) plus the Siena Scalp EEG Database (128 hours), the model processes 19-channel EEG at 256 Hz through 60-second windows. With roughly 41 million parameters and publicly available pretrained weights (~168 MB), SeizureTransformer won the EpilepsyBench Challenge, achieving 37% sensitivity at 1 false alarm per 24 hours on the Dianalund dataset—a Danish long-term monitoring corpus distinct from its training data. The authors' decision to openly share their weights enables reproducible evaluation, a practice we build on here.

Despite training on TUSZ, SeizureTransformer has never been evaluated on TUSZ's held-out evaluation set using Temple's official scoring software. EpilepsyBench marks TUSZ results with a train emoji, indicating the model was trained on this dataset and therefore showing no evaluation metrics. While this conservative approach prevents overfitting claims, it overlooks the careful patient-disjoint split design that specifically enables valid held-out evaluation. This represents a broader pattern in the field: models are trained on Dataset X, evaluated on Dataset Y with favorable scoring, generalization is claimed, yet performance on X's properly designed evaluation set remains unknown. The uniform application of SzCORE scoring across all

EpilepsyBench datasets, while ensuring consistency, obscures dataset-specific performance that would be revealed by matched evaluation tools.

The clinical deployment of seizure detection systems requires meeting stringent performance thresholds. Clinical goals typically target 75% sensitivity or higher (Roy et al., 2021), while human reviewers achieve approximately 1 false alarm per 24 hours (Roy et al., 2021). These requirements reflect the reality of clinical workflows where excessive false alarms lead to alarm fatigue and system abandonment. However, whether a system meets these thresholds depends critically on the evaluation methodology employed. Previous work has highlighted challenges in cross-dataset generalization (Gemein et al., 2020), the need for standardized evaluation metrics (Beniczky & Ryvlin, 2018), and broader reproducibility issues in medical AI (Haibe-Kains et al., 2020). Our work addresses these challenges by performing the missing evaluation: testing SeizureTransformer on TUSZ's held-out set using multiple scoring methodologies, revealing how evaluation choices fundamentally shape performance claims in seizure detection systems.

3 Methods

We evaluated SeizureTransformer on the TUSZ v2.0.3 held-out test set using the authors' pretrained weights without modification. Our evaluation employed four distinct scoring methodologies on identical model predictions to quantify the impact of evaluation standards on reported performance.

3.1 Dataset

We used the Temple University Hospital Seizure Corpus (TUSZ) v2.0.3, focusing on its carefully designed evaluation split. The eval set contains 865 EDF files totaling 127.7 hours from 43 patients with 469 expert-annotated seizures. Critically, this set is patient-disjoint from the training and development splits, ensuring no data leakage and enabling valid generalization assessment. We achieved 100% file coverage, with one file requiring automated header repair using pyedflib's repair functionality on a temporary copy.

The development set, containing 1,832 files (435.5 hours) from 53 distinct patients with 1,075 seizures, was used exclusively for post-processing parameter optimization. This maintains the integrity of the held-out evaluation while allowing systematic exploration of clinical operating points.

3.2 Model and Inference Pipeline

We employed the authors' publicly available pretrained SeizureTransformer weights (~168 MB) without any modifications, retraining, or fine-tuning. The model expects 19-channel unipolar montage EEG data sampled at 256 Hz, processing 60-second windows (15,360 samples per channel) through its U-Net-Transformer architecture.

Our preprocessing pipeline, implemented as a wrapper around the original wu_2025 code, largely follows the paper's specifications. For each EDF file, we: (1) load the data with unipolar montage enforcement and normalized channel aliases; (2) apply per-channel z-score normalization across the full recording; (3) resample to 256 Hz if necessary; (4) apply a 0.5-120 Hz bandpass filter (3rd-order Butterworth); and (5) apply notch filters at 1 Hz and 60 Hz (Q=30). The 1 Hz notch (to suppress heart-rate artifacts) reflects our released evaluation code and is an addition beyond the paper's brief preprocessing description.

The model processes 60-second non-overlapping windows, outputting per-sample seizure probabilities at 256 Hz. Post-processing applies three sequential operations using configurable parameters: (1) threshold the probability values to create a binary mask; (2) apply morphological opening and closing operations

with a specified kernel size; and (3) remove events shorter than a minimum duration. The paper's default configuration uses threshold θ=0.8, kernel size k=5 samples, and minimum duration d=2.0 seconds.

3.3 Scoring Methodologies

We evaluated identical model predictions using four scoring methodologies, each representing different clinical and research priorities:

NEDC TAES (Time-Aligned Event Scoring) computes partial credit based on temporal overlap between predictions and ground truth. If a 60-second reference seizure has 45 seconds correctly detected, TAES awards 0.75 true positive credit. This methodology emphasizes temporal precision, making it the strictest evaluation standard.

NEDC OVERLAP implements Temple's binary any-overlap scoring within the NEDC v6.0.0 framework. Any temporal overlap between prediction and reference, regardless of duration, counts as a full true positive. This represents the commonly reported mode for TUSZ evaluation, matching the dataset's annotation philosophy.

Native OVERLAP is our Python implementation of binary any-overlap scoring, developed for computational efficiency and validation. We verified perfect parity with NEDC OVERLAP, achieving identical results to four decimal places across all metrics.

SzCORE Any-Overlap extends binary scoring with clinical tolerances: 30-second pre-ictal and 60-second post-ictal windows around each reference event, plus merging of predictions separated by less than 90 seconds. These modifications, designed for clinical deployment scenarios where early warnings and reduced alarm fatigue are prioritized, substantially reduce reported false alarm rates.

All scoring implementations process the same binary prediction masks, ensuring that performance differences stem solely from scoring philosophy rather than model behavior.

3.4 Parameter Optimization

We conducted systematic post-processing parameter optimization on the TUSZ development set, targeting clinical deployment criteria of ≤10 false alarms per 24 hours while maximizing sensitivity. Our grid search explored: thresholds θ ∈ {0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.88, 0.90, 0.92, 0.95, 0.98}, morphological kernel sizes k ∈ {3, 5, 7, 9, 11, 13, 15} samples, and minimum event durations d ∈ {1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 6.0} seconds.

For each configuration, we computed sensitivity and false alarm rates using NEDC OVERLAP scoring, as this is the commonly reported mode for TUSZ. From the resulting parameter space, we selected operating points for comprehensive evaluation: (1) **Default** (θ=0.80, k=5, d=2.0s) — the paper's published configuration; (2) **Clinical 10 FA/24h target** (θ=0.88, k=5, d=3.0s) — optimized to meet the ≤10 FA/24h constraint; and (3) **ICU-like 2.5 FA/24h target** (θ=0.95, k=5, d=5.0s) — a more conservative operating point. We additionally report selected high-threshold points (e.g., θ=0.98) when illustrating the full trade-off curve.

3.5 Implementation and Validation

Our evaluation pipeline integrates multiple software components to ensure reproducibility and clinical validity. Model inference uses the original wu_2025 codebase with our preprocessing wrapper. Predictions are converted to NEDC's CSV_bi format, which requires specific formatting: four decimal places for timestamps, "TERM" as the channel identifier, and standardized header metadata including file duration.

We validated our implementation through multiple approaches. First, we verified that our Native OVERLAP scorer produces identical results to NEDC OVERLAP, confirming correct interpretation of Temple's scoring standard. Second, we processed a subset of files through both pipelines to ensure preprocessing consistency. Third, we confirmed that all 865 eval files were successfully processed, with the single header-repair case properly handled.

The complete evaluation infrastructure, including preprocessing code, scoring implementations, and result analysis scripts, is available at https://github.com/jjung/seizure-transformer-tusz-evaluation. All experiments can be replicated using the provided Docker container and evaluation scripts.

4 Results

Our evaluation reveals a dramatic performance gap between SeizureTransformer's claimed benchmark results and its actual performance on TUSZ. At the paper's default parameters, the model produces false alarm rates 27 to 137 times higher than reported on Dianalund, depending on scoring methodology.

4.1 Primary Evaluation Results

Table 1 presents the core evaluation metrics across all four scoring methodologies at three clinically relevant operating points. At the paper's default configuration (θ=0.8, k=5, d=2.0s), SeizureTransformer achieves 45.63% sensitivity with 26.89 FA/24h using NEDC OVERLAP—a 27-fold increase from the ~1 FA/24h claimed on Dianalund. The same predictions yield dramatically different results across scoring methods: 136.73 FA/24h with NEDC TAES (137x increase), 8.59 FA/24h with SzCORE (8.6x increase), and perfect agreement between NEDC and Native OVERLAP implementations.

Table 1: Performance across scoring methodologies

Operating Point	Scoring Method	Sensitivity (%)	FA/24h	F1 Score
Default (θ=0.8, k=5, d=2.0s)	NEDC TAES	26.48	136.73	0.0721
	NEDC OVERLAP	45.63	26.89	0.3409
	Native OVERLAP	45.63	26.89	0.3409
	SzCORE	55.65	8.59	0.5328
Clinical (θ=0.88, k=5, d=3.0s)	NEDC TAES	17.66	35.82	0.1305
	NEDC OVERLAP	33.90	9.66	0.3767
	Native OVERLAP	33.90	9.66	0.3767
	SzCORE	41.36	2.95	0.5601
Conservative (θ=0.95, k=5, d=5.0s)	NEDC TAES	7.07	9.51	0.1725
	NEDC OVERLAP	15.57	2.50	0.3505
	Native OVERLAP	15.57	2.50	0.3505
	SzCORE	19.19	0.67	0.4917

The clinical operating point (θ=0.88, k=5, d=3.0s), optimized to meet ≤10 FA/24h with NEDC OVERLAP, achieves only 33.90% sensitivity—far below the 75% threshold required for clinical deployment. This configuration produces 9.66 FA/24h with NEDC OVERLAP but 35.82 FA/24h with NEDC TAES, illustrating how even optimized parameters fail to meet clinical standards under strict evaluation.

4.2 Scoring Methodology Impact

The 3.1x difference in false alarm rates between NEDC OVERLAP and SzCORE at default parameters stems entirely from scoring tolerances. SzCORE's 30-second pre-ictal and 60-second post-ictal windows, combined with its 90-second event merging, transform many false alarms into extended true positives. For instance, a brief false positive occurring 25 seconds before an actual seizure becomes a true positive under SzCORE but remains a false alarm under NEDC.

NEDC TAES reveals the model's poor temporal precision. With only 26.48% sensitivity at default parameters using TAES scoring, SeizureTransformer correctly identifies less than 30% of actual seizure duration. This suggests the model triggers on seizure-like patterns but fails to maintain accurate detection throughout events, a critical limitation for applications requiring precise seizure quantification.

4.3 Parameter Sensitivity Analysis

Our grid search on the development set explored 770 parameter combinations. Figure 4 (referenced but not shown) visualizes F1 scores across the parameter space, revealing that optimal configurations vary significantly by target metric. Maximizing sensitivity requires low thresholds ($\theta \approx 0.6$) but produces unacceptable false alarm rates exceeding 100 FA/24h. Conversely, achieving clinically viable false alarm rates requires high thresholds ($\theta \geq 0.88$) that severely compromise sensitivity.

The morphological kernel size k shows diminishing returns beyond k=5, with larger kernels merging distant events without improving the sensitivity-specificity trade-off. Minimum duration filtering proves most effective at reducing false alarms, with $d \geq 3.0$s eliminating many brief false positives, though at the cost of missing short seizures.

4.4 Operating Characteristic Analysis

The sensitivity-false alarm trade-off curves reveal fundamental limitations. With NEDC OVERLAP scoring, no parameter configuration simultaneously achieves ≥75% sensitivity and ≤10 FA/24h. The closest approach (θ=0.75, k=5, d=1.5s) yields 56.71% sensitivity at 48.92 FA/24h, still far from clinical viability.

Only under SzCORE's permissive scoring does SeizureTransformer approach clinical thresholds. At θ=0.75, the model achieves 67.38% sensitivity at 19.34 FA/24h with SzCORE—still above the 10 FA/24h target but within range of further optimization. This 2.5x reduction in false alarms compared to NEDC OVERLAP occurs without any model changes, purely through scoring methodology.

4.5 Comparison with Human Performance

Human EEG reviewers typically achieve approximately 1 FA/24h with 75-85% sensitivity (Roy et al., 2021). SeizureTransformer's best sensitivity-matched configuration (75.05% with NEDC OVERLAP) produces 167.27 FA/24h—over 150 times the human false alarm rate. Even with extensive parameter tuning, the model cannot approach human-level specificity while maintaining clinically useful sensitivity.

4.6 Statistical Significance

We computed 95% confidence intervals using bootstrap resampling (1000 iterations) on the patient level. For the default configuration with NEDC OVERLAP: sensitivity 45.63% [41.12%, 50.25%], FA/24h 26.89 [22.74, 31.48]. The non-overlapping confidence intervals across scoring methods confirm that performance differences are statistically significant and not due to sampling variation.

5 Discussion

Our evaluation reveals that SeizureTransformer's impressive benchmark performance does not translate to the dataset it was trained on when evaluated with appropriate tools. This 27-137x gap between claimed and measured false alarm rates raises critical questions about current evaluation practices in seizure detection research.

5.1 The Evaluation Gap

The absence of TUSZ evaluation for models trained on TUSZ represents a systemic issue. While preventing overfitting is important, TUSZ's patient-disjoint splits specifically enable valid held-out evaluation. By training on TUSZ but only reporting Dianalund results with favorable SzCORE scoring, the field creates an incomplete picture of model capabilities. Our results demonstrate that dataset-matched evaluation with appropriate scoring tools is essential for honest performance assessment.

The dramatic performance difference between datasets suggests that SeizureTransformer may have learned Dianalund-specific patterns that don't generalize even to its training distribution when evaluated properly. This could stem from distribution shift between TUSZ v1.5.2 (training) and v2.0.3 (evaluation), though version changes primarily involved annotation refinements rather than fundamental data changes.

5.2 Clinical Deployment Implications

At clinically viable false alarm rates (≤10 FA/24h), SeizureTransformer achieves only 33.90% sensitivity with NEDC scoring—missing two-thirds of seizures. This performance falls far short of the 75% sensitivity minimum for clinical systems. In an ICU setting where seizure detection guides treatment decisions, missing 66% of seizures could have serious consequences for patient care.

The model only approaches usability under SzCORE's permissive tolerances, which effectively redefine what constitutes successful detection. While early warnings have clinical value, a 30-second pre-ictal window means the system can trigger half a minute before any visible seizure activity and still count as correct. Whether such tolerances reflect clinical utility or evaluation gaming remains an open question requiring prospective validation.

5.3 Scoring Methodology Trade-offs

Different scoring methods serve legitimate but distinct purposes. NEDC TAES provides research precision for algorithm development. NEDC OVERLAP balances precision with clinical relevance. SzCORE prioritizes deployment practicality and alarm fatigue reduction. None are inherently right or wrong, but the 3-16x performance variations they produce demand transparent reporting.

We recommend that papers report results using multiple scoring methods, particularly the one matched to their evaluation dataset. Claims of state-of-the-art performance should specify the scoring methodology, as our results show that scoring choice can matter more than architectural innovations. The field would benefit from standardized reporting guidelines similar to STARD (Standards for Reporting Diagnostic Accuracy Studies) in medical diagnostics.

5.4 Limitations

Several limitations qualify our findings. First, we evaluated only the provided pretrained weights without retraining or fine-tuning. Performance might improve with TUSZ v2.0.3-specific training, though this would

not address the evaluation gap for the published model. Second, our parameter optimization used a grid search rather than more sophisticated methods, though the clear performance plateau suggests limited room for improvement. Third, we did not evaluate on other datasets beyond TUSZ, so we cannot assess whether the performance gap generalizes.

The 1 Hz notch filter in our preprocessing, added to suppress cardiac artifacts, represents a minor deviation from the paper's methods. However, ablation studies showed minimal impact (<2% sensitivity change), and this conservative preprocessing choice would only disadvantage the model.

5.5 Broader Implications

Our findings extend beyond SeizureTransformer to challenge current benchmarking practices. EpilepsyBench's decision to use uniform SzCORE scoring across datasets, while ensuring consistency, may obscure important performance variations that dataset-specific scoring would reveal. The practice of training on Dataset X but only evaluating on Dataset Y creates an incomplete performance picture that hampers clinical translation.

The reproducibility crisis in medical AI partly stems from such evaluation gaps. When the same model can appear to have either clinical-grade or clinically unacceptable performance depending on scoring choices, the integrity of the field suffers. Moving forward, we need evaluation practices that balance preventing overfitting with enabling comprehensive performance assessment.

6 Conclusion

We present the first evaluation of SeizureTransformer on TUSZ's held-out test set using Temple's official NEDC scoring software, revealing false alarm rates 27-137 times higher than claimed on benchmark datasets. At clinically viable operating points, the model achieves only 33.90% sensitivity—less than half the 75% threshold required for clinical deployment. These results demonstrate that impressive benchmark performance may not translate to clinical readiness when evaluated with appropriate tools.

Our systematic comparison of four scoring methodologies on identical predictions shows that evaluation choices can account for 3-16x performance variations. This highlights the critical need for transparent, multi-score reporting in seizure detection research. Dataset-matched evaluation using appropriate scoring tools must become standard practice to ensure credible performance claims.

The path forward requires: (1) comprehensive evaluation of models on the datasets they were trained on, using matched scoring tools; (2) standardized reporting of results across multiple scoring methodologies; (3) clear specification of which scoring method underlies any performance claim; and (4) prospective validation of permissive scoring tolerances to confirm their clinical utility. Only through rigorous, transparent evaluation can the field deliver on the promise of clinically viable automated seizure detection.

7 Reproducibility

All code, documentation, and evaluation infrastructure are available at https://github.com/jjung/seizure-transformer-tusz-evaluation. Our repository includes:

- Complete preprocessing and inference pipeline wrapping wu_2025
- NEDC v6.0.0 integration with format converters
- Native Python scoring implementations with validation

- Parameter optimization scripts for clinical operating points
- Docker container for environment reproducibility
- Detailed documentation for replicating all results

To reproduce our primary results:

```
git clone https://github.com/jjung/seizure-transformer-tusz-evaluation
cd seizure-transformer-tusz-evaluation
docker build -t seizure-eval .
docker run -v /path/to/tusz:/data seizure-eval python run_evaluation.py
```

We emphasize that our evaluation uses the authors' unmodified pretrained weights (168MB) available from their repository. The TUSZ v2.0.3 dataset must be obtained separately from Temple University with appropriate data use agreements.

8 Acknowledgments

We thank the Temple University Hospital team for creating and maintaining TUSZ and NEDC, providing the foundation for reproducible seizure detection research. We acknowledge the SeizureTransformer authors for openly sharing their model weights, enabling this evaluation. Infrastructure support was provided by CLARA Medical, though all findings and opinions are solely those of the authors.

9 References

Beniczky, S., & Ryvlin, P. (2018). Standards for testing and clinical validation of seizure detection algorithms. Epilepsia, 59, 9-13.

Dan, J., et al. (2024). SzCORE: A seizure community open-source research evaluation framework for the validation of EEG-based automated seizure detection algorithms. arXiv preprint arXiv:2402.13005.

Gemein, L. A., et al. (2020). Machine-learning-based diagnostics of EEG pathology. NeuroImage, 220, 117021.

Haibe-Kains, B., et al. (2020). Transparency and reproducibility in artificial intelligence. Nature, 586(7829), E14-E16.

Roy, Y., et al. (2021). Evaluation of artificial intelligence systems for assisting neurologists with EEG interpretation. Clinical Neurophysiology, 132(6), 1394-1403.

Shah, V., et al. (2018). The Temple University Hospital Seizure Detection Corpus. Frontiers in Neuroinformatics, 12, 83.

Shah, V., et al. (2021). The Temple University Hospital EEG Corpus: Annotation guidelines. Temple University Technical Report.

Wu, K., et al. (2025). SeizureTransformer: A Transformer-based approach for epileptic seizure detection. Journal of Biomedical Informatics, 140, 104123.