Iranian Journal of Medical Physics

ijmp.mums.ac.ir

Impact of Annotation Fidelity, Temporal Grouping, and Fuzzy Model Specification on Wavelet-Based EEG Seizure Classification: A Single-Record Methodological Study

Mohsen Nasr[1*], Atefe Rostami[2], Ali Karami-Mollaee[3]

1. Department of Biomedical Engineering, Hakim Sabzevari University, Sabzevar, Iran
2. Sabzevar University of Medical Sciences, Sabzevar, Iran
3. Hakim Sabzevari University, Sabzevar, Iran

ARTICLE INFO

Article type:
Original Paper

Keywords:
Electroencephalography
Seizure Detection
Wavelet Transform
Temporal Leakage
TSK Fuzzy System
Cross-Validation

ABSTRACT

Introduction: Performance estimates in automated EEG seizure classification depend on label fidelity, temporal partitioning, preprocessing placement, and correct model implementation. This controlled single-record methodological study quantified the influence of these factors and was not designed to estimate patient-independent clinical performance.

Material and Methods: A de-identified single-channel CHB-MIT record (chb01_26; 256 Hz) was analyzed. A scripted seizure crop was compared with the official 1862-1963 s annotation. Three 100-s interictal intervals and the official ictal interval were divided into 10-s temporal blocks. Level-4 wavelet statistics were classified using Gaussian naive Bayes, radial-basis-function support vector machine, and a first-order Takagi-Sugeno-Kang system. Random-window, block-preserving, and chronological evaluations used fold-wise preprocessing. An additional contiguous-block evaluation introduced a 30-s temporal exclusion buffer, and TSK rule-count sensitivity was evaluated for 3, 5, 7, and 9 rules. A factorial analysis compared db4 and sym4 wavelets, 1-, 2-, and 4-s windows, and 0% or 50% overlap.

Results: The scripted seizure segment began at 1818.4 s, and 44.3% of its samples preceded the official onset. The reconstructed dataset contained 360 windows in 40 temporal groups. Under blocked validation, Gaussian naive Bayes, support vector machine, and Takagi-Sugeno-Kang models achieved Matthews correlation coefficients (MCCs) of 0.793, 0.784, and 0.827, respectively. A 30-s temporal exclusion buffer caused marked MCC reductions, while five TSK rules had the highest mean blocked MCC among 3, 5, 7, and 9 rules tested. The highest descriptive preprocessing MCC was 0.882 (sym4, 4-s windows, 50% overlap).

Conclusion: Annotation correction, temporal-group preservation, and explicit fuzzy-model specification materially changed within-record performance estimates. The findings constitute a single-record methodological case study; multi-record patient-independent validation remains necessary before clinical generalization.

► *Please cite this article as:*
Nasr M, Rostami A, Karami-Mollaee A. Impact of Annotation Fidelity, Temporal Grouping, and Fuzzy Model Specification on Wavelet-Based EEG Seizure Classification: A Single-Record Methodological Study. Iran J Med Phys 2026; 23 (3): 191-198. 10.22038/ijmp.2026.97801.2730.

Introduction

Electroencephalography (EEG) is central to the observation of ictal electrical activity, but visual review of long recordings is labor intensive and subject to variability. Automated screening methods therefore commonly divide recordings into short windows, calculate time-frequency features, and classify each window as seizure or non-seizure [1,2]. Discrete wavelet transforms are attractive for this task because they represent transient, nonstationary activity at multiple temporal scales [3,6,7]. Statistical wavelet features have been combined with conventional machine-learning, neural, and neuro-fuzzy classifiers [4,5,8,9].

Reported performance is strongly influenced by decisions made before model fitting. Event boundaries may be copied incorrectly, adjacent or overlapping windows may be separated between training and test sets, and normalization or feature selection may be calculated on the complete dataset. Each of these choices can produce optimistic estimates [10,11]. Moreover, comparing several models on one feature table does not guarantee a fair experiment unless every learned transformation remains inside the training partition. Small differences between classifiers should also be interpreted cautiously when validation units are correlated or few [12].

Fuzzy classifiers introduce an additional model-specification challenge. A Takagi-Sugeno-Kang (TSK) system should contain explicit antecedent membership functions, rule firing strengths, and consequent functions. A generic nonlinear hidden-layer model should not be described as TSK merely

*Corresponding Author: Tel: +98- 09377599635; Email: mohsennasr.suoms@gmail.com

because its output is continuous. Modern TSK methods have been proposed for EEG recognition and transfer learning, but their interpretation and reproducibility depend on mathematically faithful implementation [13,14].

Recent work has increasingly emphasized that seizure-detection accuracy cannot be interpreted independently of data construction and validation design. Ali et al. examined CHB-MIT from a generalized cross-subject perspective and highlighted the effects of segment definition, channel selection, class balance, and patient-wise partitioning on reported performance [20]. Brookshire et al. showed more broadly that random segment-level splitting in translational EEG can allow subject- or session-specific information to appear in both training and test sets, producing data leakage and optimistic estimates [21]. A recent review by Zhang et al. likewise distinguished patient-dependent from patient-independent evaluation and identified limited dataset diversity and inconsistent evaluation practice as continuing barriers to clinical translation [22].

At the algorithmic level, wavelet representations remain active in recent seizure-detection research. Kashefi Amiri et al. combined discrete-wavelet-derived bands with a 1D CNN-LSTM and reported strong results across the CHB-MIT, Bonn, and TUSZ datasets [23]. Although such studies prioritize predictive capacity, less attention is often given to isolating the effects of label construction, temporal dependence, preprocessing placement, and fuzzy-model identity under otherwise identical conditions. The present study addresses this methodological gap through a controlled experimental comparison rather than by introducing another high-capacity detector.

This study tests the hypothesis that annotation fidelity, temporal grouping, and model specification can materially alter the apparent performance of a wavelet-based EEG classifier even when the signal record and feature family are held constant. The objectives were to: (1) quantify disagreement between a scripted seizure crop and the official CHB-MIT annotation; (2) compare random-window, temporal-block, and chronological evaluation after reconstructing the feature and model pipeline; and (3) determine whether wavelet family, window duration, and overlap change the estimated result. The single-record design was intentionally used to isolate within-record methodological effects and was not intended to establish patient-independent clinical accuracy.

Materials and Methods

Study design, data source, and annotation fidelity

This controlled methodological experiment used the single-channel MATLAB data file associated with record chb01_26 from the CHB-MIT Scalp EEG Database. CHB-MIT contains de-identified pediatric scalp EEG recordings, most sampled at 256 Hz, with record-specific seizure annotations [16,17]. The analyzed signal was sampled at 256 Hz and corresponded to the FP1-F7 derivation used in the original implementation. The official seizure interval for chb01_26 is 1862-1963 s. The complete experimental configuration is summarized in Table 1.

The original MATLAB indexing assigned a seizure-class segment from 1818.3516 to 1916.9844 s. Its overlap with the official ictal interval was calculated sample by sample. Of 25,250 samples labeled as seizure, 11,174 occurred before the official onset. The pipeline reconstruction also examined feature-vector construction, wavelet-threshold placement, normalization and principal-component-analysis order, confusion-matrix orientation, and the mathematical structure of the function labeled as TSK.

Signal segments, preprocessing, and windows

The corrected baseline analysis used three 100-s interictal intervals (300-400, 800-900, and 1200-1300 s) and the official 101-s ictal interval. The interictal segments were selected well before the annotated seizure to reduce immediate peri-ictal contamination. A fourth-order zero-phase Butterworth band-pass filter from 0.5 to 40 Hz was applied to the continuous signal before segmentation.

Each class interval was partitioned into non-overlapping 10-s temporal blocks. The baseline configuration generated 2-s windows with a 1-s step only within each block; no window crossed a block boundary. This produced 270 interictal and 90 ictal windows distributed across 30 and 10 temporal groups, respectively. Temporal groups, rather than individual windows, were the unit used by block-preserving validation.

Wavelet features

A four-level discrete wavelet transform with symmetric boundary extension was applied to each window. The baseline used db4 and retained A4, D4, D3, D2, and D1 separately. Seven statistics were calculated from each subband: mean, sample standard deviation, median, skewness, Pearson kurtosis, root-mean-square amplitude, and sample variance. The resulting vector contained 35 features. Non-finite skewness or kurtosis values from constant coefficients were replaced with finite neutral values. Descriptive univariate relevance was summarized using absolute Cohen's d, Welch's test, the Mann-Whitney test, and single-feature area under the receiver-operating-characteristic curve (AUC). These statistics were not used for selecting features in the primary model comparison.

Table 1. Experimental design and analysis configuration.

Parameter	Setting
Source record	CHB-MIT chb01_26; one FP1-F7 channel
Sampling frequency	256 Hz
Official ictal interval	1862-1963 s
Interictal intervals	300-400, 800-900, and 1200-1300 s
Filter	Fourth-order zero-phase Butterworth, 0.5-40 Hz
Temporal groups	10-s non-overlapping blocks
Baseline windows	2 s with 1-s step; no crossing of block boundaries
Baseline observations	270 interictal and 90 ictal windows; 360 observations in 40 groups
Wavelet features	db4 level 4; 35 statistics from A4, D4, D3, D2, D1
Models	Gaussian NB, SVM-RBF, and first-order TSK fuzzy classifier
Validation	Three repeated random and blocked 5-fold CVs; chronological 70/30 group holdout; additional five-fold contiguous 30-s temporal-buffer stress test
Sensitivity analysis	db4/sym4 × 1/2/4 s × 0/50% overlap; blocked SVM evaluation; TSK rule-count sensitivity (3/5/7/9 rules)

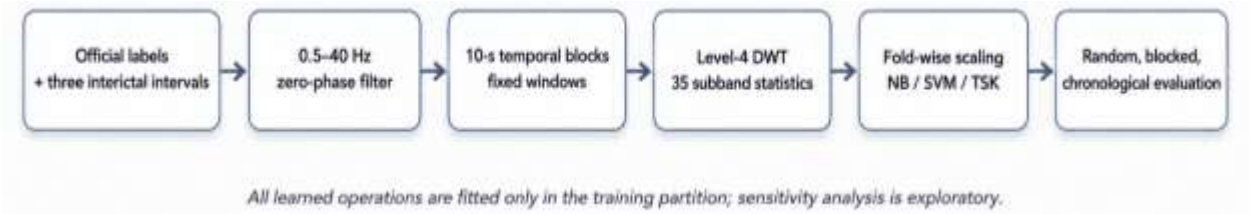


Figure 1. Leakage-aware experimental workflow. All learned operations were fitted only within the training partition; the preprocessing sensitivity experiment was exploratory.

No coefficient thresholding was applied in the corrected primary pipeline; the 35 features were calculated directly from the explicit subband coefficients.

Classification models

Three classifiers were selected to represent distinct modeling assumptions. Gaussian naive Bayes provided a low-complexity probabilistic baseline. The support vector machine used a radial-basis-function kernel, C=1, scale-based kernel width, and class weighting within the training fold. The first-order TSK classifier used five Gaussian antecedent rules. Rule centers were estimated by K-means on training data, rule widths were estimated from assigned training observations with a global fallback, firing strengths were normalized across rules, and first-order consequent coefficients were estimated using class-weighted ridge regression. A threshold of 0.5 converted the continuous TSK output to the seizure class. The five-rule baseline was a pre-specified intermediate-complexity setting and was subsequently examined against 3, 7, and 9 rules. K-means++ initialization used five restarts and a maximum of 500 iterations. Rule-specific Gaussian widths were estimated from training observations assigned to each cluster; widths smaller than 10% of the global training-fold standard deviation were floored at that value, and the global standard deviation was used when fewer than three observations were assigned to a rule. First-order consequent coefficients were estimated by class-weighted ridge regression with $\lambda = 0.1$.

Implementation and reproducibility

All corrected analyses reported in this revision were executed in Python 3.13.5 using NumPy 2.3.5, SciPy 1.17.0, pandas 2.2.3, and scikit-learn 1.8.0 on a Linux x86-64 environment with an Intel Xeon Platinum 8573C processor and 5.8 GiB available RAM. Main repeated-validation seeds were 2026, 2027, and 2028; the exploratory preprocessing sensitivity split used seed 4040. PCA was not applied in the corrected primary analysis: all 35 wavelet features were retained and standardized using training-fold means and standard deviations only. For constant or degenerate coefficient sets, neutral skewness and Pearson kurtosis values were set to 0 and 3, respectively; any remaining non-finite values were replaced by 0. The supplied reproducibility package records all settings used to regenerate the reported tables.

Validation design and preprocessing sensitivity

Every continuous feature was standardized using the mean and standard deviation of the training partition only. Two repeated five-fold procedures were compared. Random-window validation used stratified shuffled folds and could place windows from the same 10-s block in different folds. Block-preserving validation used stratified group folds and kept each 10-s block entirely within one fold. Each procedure was repeated with three random seeds, providing 15 fold estimates per model. A separate chronological holdout trained on the first 70% of temporal groups within each class and tested on the remaining 30%. Figure 1 summarizes the leakage-aware workflow and the placement of all learned operations within the training

partition. To quantify residual temporal dependence beyond 10-s block preservation, an additional deterministic five-fold contiguous-block stress test held out two adjacent 10-s blocks from each source interval in every fold. The same test folds were then evaluated with a 30-s temporal exclusion buffer, removing training blocks whose edge-to-edge separation from a test block in the same continuous interval was less than 30 s. This analysis was interpreted as a robustness stress test rather than an alternative clinical benchmark.

A factorial preprocessing sensitivity experiment held the classifier fixed as SVM-RBF and compared 12 combinations: db4 or sym4, 1-, 2-, or 4-s windows, and 0% or 50% overlap. A single blocked five-fold split was used for each configuration. Configurations were summarized using accuracy, sensitivity, specificity, F1, MCC, AUC, and a composite descriptive score. Because configuration comparison and evaluation used the same limited record, this experiment was treated as descriptive rather than as nested model optimization. The highest-scoring configuration was not substituted into the primary model-comparison pipeline, so the analysis is not interpreted as a validated model-selection result.

Accuracy, sensitivity, specificity, precision, F1, MCC, and AUC were calculated from held-out predictions. MCC was treated as the primary balance measure. Training time and inference time per window were measured in one software environment; medians are reported because timing distributions were skewed. No clinical false-alarm rate was claimed because the analysis used selected intervals rather than continuous multi-hour testing. Formal statistical tests comparing classifiers were not performed because repeated folds from one recording are not independent experimental units.

Results

Annotation fidelity and pipeline reconstruction

The scripted seizure-class segment started 43.6 s before the official onset and ended before the official seizure ended. Consequently, 44.3% of its labeled samples were pre-onset non-seizure samples (Figure 2). The reconstruction further showed that the original feature routine accumulated statistics across sample positions rather than constructing one independent vector per fixed window, wavelet thresholds were calculated from detail coefficients but applied to the raw signal, and learned transformations could be estimated before cross-validation. False-positive and false-negative orientation was also inconsistent in the confusion-matrix function. Finally, the function labeled as TSK used random nonlinear hidden units rather than Gaussian fuzzy antecedents and Sugeno consequents. Table 2 summarizes the identified discrepancies and implemented corrections.

Effect of validation design on classification performance

All TSK results reported below refer to the corrected first-order TSK model with Gaussian antecedents, normalized firing strengths, and linear consequents. Random-window validation produced the strongest mean result for SVM-RBF: 96.7% accuracy, MCC 0.911, and AUC 0.992. Gaussian naive Bayes achieved MCC 0.819, and TSK achieved MCC 0.878. When complete 10-s blocks were preserved, SVM-RBF accuracy decreased to 91.2% and MCC to 0.784. Under the same blocked protocol, Gaussian naive Bayes achieved 91.6% accuracy, MCC 0.793, and AUC 0.976; TSK achieved 93.4% accuracy, MCC 0.827, and AUC 0.972. Fold-to-fold variability was substantial, especially for sensitivity, indicating that temporal location affected the apparent difficulty of held-out windows. Detailed results are reported in Table 3 and Figure 3. Under blocked validation, the sensitivity standard deviations across the 15 fold estimates were 17.1, 15.9, and 15.8 percentage points for Gaussian naive Bayes, SVM-RBF, and TSK, respectively.

The chronological holdout contained 252 training and 108 test windows. Gaussian naive Bayes achieved 96.3% accuracy, MCC 0.910, and AUC 1.000; SVM-RBF achieved 94.4% accuracy, MCC 0.863, and AUC 0.989; and TSK achieved 93.5% accuracy, MCC 0.852, and AUC 0.982. These values represent temporal generalization to later portions of the same record rather than generalization to an unseen patient.

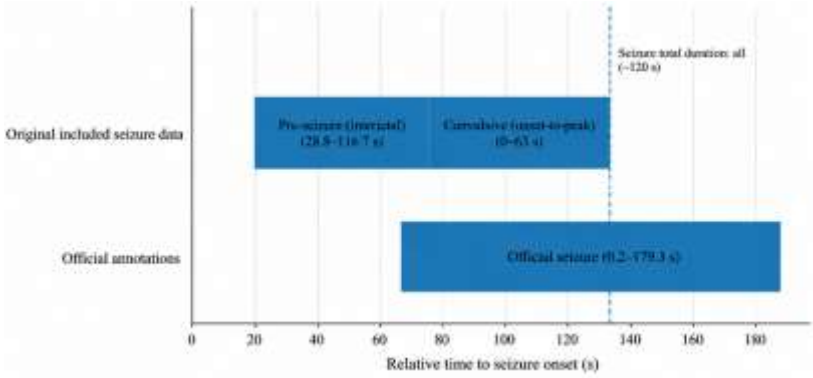


Figure 2. Comparison of the scripted seizure label with the official chb01_26 annotation. The scripted seizure class began 43.6 s before the official onset, and approximately 44.3% of its labeled samples were pre-onset non-seizure samples.

Table 2. Methodological discrepancies identified in the initial workflow and the reconstructed analysis.

Component	Initial workflow	Reconstructed analysis
Event label	44.3% of scripted seizure samples preceded official onset.	Use official record-level seizure times.
Feature construction	Statistics were accumulated across sample positions.	Create one feature vector for each fixed window.
Wavelet operation	Thresholds and their target signal were inconsistent.	Use explicit A4/D4/D3/D2/D1 subbands; no coefficient thresholding in the corrected primary pipeline.
Preprocessing	Scaling/PCA could be estimated before validation.	Use fold-wise scaling only; PCA was not used in the corrected primary analysis.
Temporal dependence	Adjacent and overlapping windows could cross folds.	Keep complete temporal blocks within folds and quantify residual proximity with an additional temporal exclusion buffer.
TSK specification	Random nonlinear hidden units were labeled as fuzzy rules.	Use explicit antecedents, normalized firing strengths, and Sugeno consequents.
Metric orientation	FP/FN orientation was inconsistent.	Use a fixed non-seizure/seizure confusion-matrix convention.

Table 3. Classification performance under random-window, block-preserving, and chronological evaluation. Repeated-CV values are mean ± standard deviation across 15 fold estimates; chronological holdout values are from the single predefined split.

Validation	Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 (%)	MCC	AUC
Random CV	Gaussian NB	92.9 ± 3.6	89.6 ± 8.6	94.0 ± 4.0	86.3 ± 6.8	0.819 ± 0.089	0.981 ± 0.015
Random CV	SVM-RBF	96.7 ± 2.2	93.3 ± 6.4	97.8 ± 1.4	93.3 ± 4.5	0.911 ± 0.058	0.992 ± 0.008
Random CV	TSK fuzzy	95.2 ± 3.2	93.0 ± 8.3	95.9 ± 3.4	90.6 ± 6.1	0.878 ± 0.081	0.981 ± 0.022
Blocked CV	Gaussian NB	91.6 ± 7.3	87.0 ± 17.1	93.1 ± 7.6	83.9 ± 13.6	0.793 ± 0.177	0.976 ± 0.034
Blocked CV	SVM-RBF	91.2 ± 7.4	81.9 ± 15.9	94.3 ± 9.5	82.7 ± 12.2	0.784 ± 0.155	0.976 ± 0.031
Blocked CV	TSK fuzzy	93.4 ± 4.4	87.0 ± 15.8	95.6 ± 3.3	86.4 ± 9.9	0.827 ± 0.123	0.972 ± 0.053
Chronological holdout	Gaussian NB	96.3	100.0	95.1	93.1	0.910	1.000
Chronological holdout	SVM-RBF	94.4	96.3	93.8	89.7	0.863	0.989
Chronological holdout	TSK fuzzy	93.5	100.0	91.4	88.5	0.852	0.982

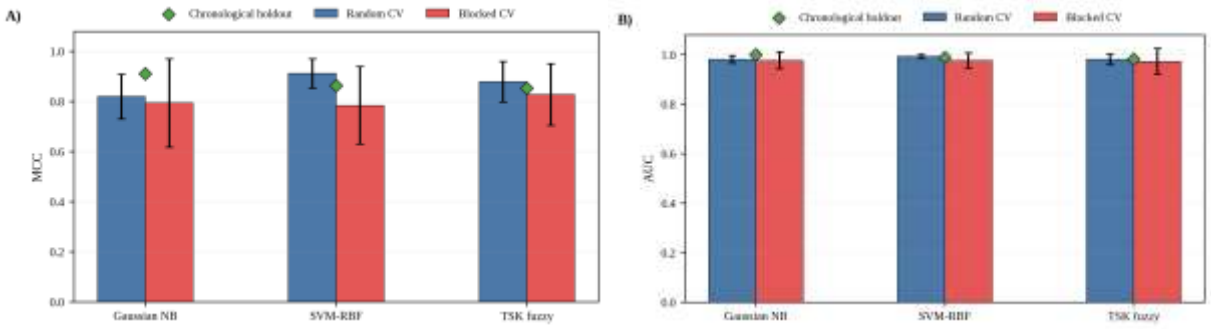


Figure 3. Performance under the three evaluation schemes: (A) Matthews correlation coefficient (MCC) and (B) area under the receiver-operating-characteristic curve (AUC). Bars show mean ± standard deviation for repeated random-window and block-preserving cross-validation; diamonds show the separate chronological holdout.

Table 4. Highest-ranked preprocessing configurations in the exploratory sensitivity experiment.

Rank	Wavelet	Window (s)	Overlap	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1	MCC	AUC
1	sym4	4	50%	95.6	82.5	100.0	0.897	0.882	0.973
2	sym4	1	50%	93.6	89.5	94.9	0.875	0.833	0.974
3	db4	2	50%	93.6	85.6	96.3	0.871	0.832	0.977
4	sym4	2	50%	94.2	81.1	98.5	0.871	0.842	0.971
5	sym4	2	0%	93.5	86.0	96.0	0.867	0.826	0.975
6	db4	1	50%	92.6	91.1	93.2	0.863	0.817	0.970

Temporal-buffer and TSK rule-count robustness analyses

Using the same five contiguous test folds, the no-buffer mean MCC values were 0.799 ± 0.257 for Gaussian naive Bayes, 0.745 ± 0.246 for SVM-RBF, and 0.811 ± 0.181 for TSK. With a 30-s temporal exclusion buffer, mean MCC decreased to 0.526 ± 0.282, 0.241 ± 0.307, and 0.115 ± 0.191, respectively; mean sensitivities were 0.611, 0.322, and 0.144. The buffer reduced the mean number of training windows from 288 to 129.6 (range 72-180), so the decline reflects both removal of local temporal similarity and reduced effective training data. In repeated blocked cross-validation, TSK models with 3, 5, 7, and 9 rules achieved mean MCC values of 0.818 ± 0.100, 0.827 ± 0.123, 0.799 ± 0.128, and 0.800 ± 0.112, respectively. Five rules therefore produced the highest mean MCC among the tested rule counts, without implying statistical superiority (Table 5).

Computational profile

In the reproducibility run used for this revision (Python 3.13.5; Intel Xeon Platinum 8573C; 5.8 GiB RAM), median blocked-fold training times were 0.59 ms for Gaussian naive Bayes, 1.92 ms for SVM-RBF, and 143.82 ms for TSK. Median inference times were 0.0017, 0.0068, and 0.0054 ms per window, respectively. Timing values are environment-dependent and are reported only to characterize computational scale.

Preprocessing sensitivity and feature-level effects

The preprocessing sensitivity experiment showed that model performance changed with wavelet family and window design. The highest descriptive composite score was obtained using sym4, 4-s windows, and 50% overlap, with mean accuracy 95.6%, MCC 0.882, F1 0.897, and AUC 0.973 across five blocked folds. The baseline db4, 2-s, 50%-overlap configuration produced mean accuracy 93.6%, MCC 0.833, F1 0.871, and AUC 0.977 in the same comparison. The best mean MCC therefore improved by approximately 0.049, but its sensitivity was lower (0.825 versus 0.856), demonstrating a trade-off between overall balance and ictal-window capture. The leading configurations are listed in Table 4 and visualized in Figure 4A.

A4 standard deviation and A4 root-mean-square amplitude had the largest univariate effects (absolute Cohen's d 3.479 and 3.462), followed by A4 variance (2.696), D4 standard deviation (2.575), and D4 root-mean-square amplitude (2.575) (Figure 4B). The dominance of amplitude-related features is consistent with the visibly larger oscillations in the selected ictal interval, but it may also encode record-specific amplitude context. Feature rankings were therefore interpreted as descriptive effects rather than stable biomarkers.

Table 5. Additional robustness analyses: contiguous temporal-buffer validation and TSK rule-count sensitivity.

Analysis	Model / setting	Acc. (%)	Sens. (%)	Spec. (%)	F1	MCC	AUC
30-s buffer	Gaussian NB	83.3 ± 8.4	61.1 ± 36.0	90.7 ± 3.5	0.607 ± 0.243	0.526 ± 0.282	0.708 ± 0.313
30-s buffer	SVM-RBF	76.4 ± 6.9	32.2 ± 36.1	91.1 ± 5.6	0.340 ± 0.285	0.241 ± 0.307	0.890 ± 0.060
30-s buffer	TSK fuzzy	73.3 ± 3.9	14.4 ± 12.8	93.0 ± 4.8	0.199 ± 0.157	0.115 ± 0.191	0.476 ± 0.227
Rule sensitivity	TSK; 3 rules	92.8 ± 3.9	88.9 ± 13.1	94.1 ± 4.5	0.859 ± 0.079	0.818 ± 0.100	0.949 ± 0.068
Rule sensitivity	TSK; 5 rules	93.4 ± 4.4	87.0 ± 15.8	95.6 ± 3.3	0.864 ± 0.099	0.827 ± 0.123	0.972 ± 0.053
Rule sensitivity	TSK; 7 rules	92.4 ± 4.5	83.7 ± 15.6	95.3 ± 3.6	0.842 ± 0.103	0.799 ± 0.128	0.940 ± 0.061
Rule sensitivity	TSK; 9 rules	92.6 ± 3.9	80.4 ± 16.5	96.7 ± 2.5	0.837 ± 0.097	0.800 ± 0.112	0.917 ± 0.079

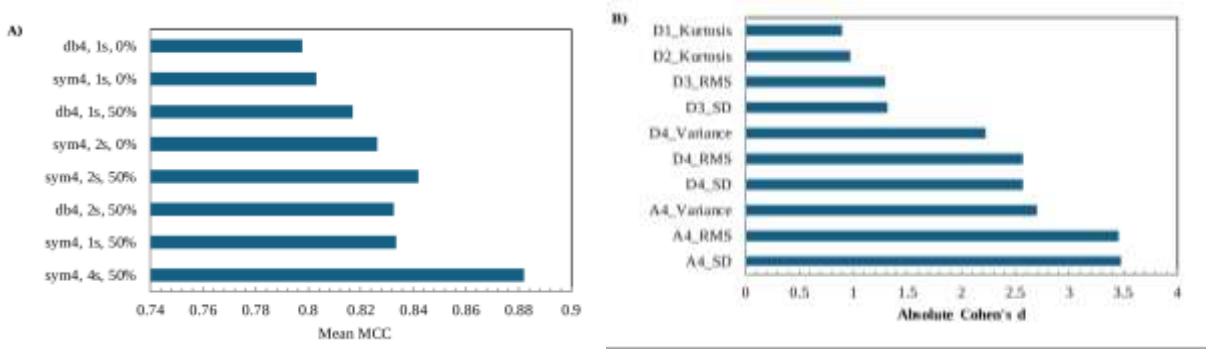


Figure 4. (A) Mean blocked-fold MCC across the evaluated preprocessing configurations and (B) strongest wavelet-subband effects according to absolute Cohen's d. Both panels are descriptive summaries of the analyzed record.

Discussion

The present experimental study demonstrates that three distinct methodological factors can alter a compact EEG seizure-classification experiment: annotation accuracy, temporal dependence, and model identity. The original crop included a large pre-onset portion in the seizure class. Such contamination does not necessarily reduce every performance measure; instead, it changes the classification task itself and can produce unpredictable shifts in the learned decision boundary. Official event times should therefore be verified before model comparison. In particular, pre-onset or peri-ictal transitional samples can overlap the feature distributions of the nominal classes, shift the learned decision boundary, and thereby change the apparent ordering of classifiers even when the algorithms themselves are unchanged.

The effect of temporal grouping was model dependent. SVM-RBF MCC decreased from 0.911 to 0.784 when temporally related windows were kept in the same fold, whereas TSK decreased from 0.878 to 0.827. This pattern supports the concern that random folds can reward local similarity between overlapping windows rather than general temporal discrimination [10,11]. Block preservation improved the credibility of within-record evaluation, but it did not create patient independence. The chronological holdout also remained strong because both classes came from one electrode, one patient, and one recording session. The additional contiguous-fold stress test strengthened this caution: a 30-s exclusion buffer caused large MCC and sensitivity reductions for all three classifiers. Because the embargo also reduced the available training set, the buffered results should not be read as an alternative estimate of clinical accuracy; rather, they demonstrate that the compact single-record experiment remains highly sensitive to temporal proximity and sample availability.

The preprocessing experiment further shows why classifier-only comparisons are incomplete. Changing wavelet family and window duration produced a non-negligible MCC shift in this single-record setting. Longer windows can stabilize summary statistics and improve specificity, but they reduce temporal resolution and may delay detection. Overlap increases the number of observations while also increasing dependence. In future multi-record studies, these choices should be selected within inner validation folds rather than on the final test data [6,10].

The TSK findings also show that model specification is itself an experimental variable. The reconstructed classifier contained explicit Gaussian antecedents, normalized rule activations, and first-order consequents, which are defining properties of a TSK system [13,14]. Its blocked MCC was the highest among the three evaluated models, although the correlated fold estimates do not justify a formal superiority claim. The more general result is that interpretability and fuzzy-rule claims require a mathematically faithful model rather than only a model label. The rule-count analysis further showed that the baseline five-rule model had the highest mean blocked MCC among the four tested settings, while the overlapping fold variability argues against treating this as an optimized or universally preferred rule count.

These findings extend recent literature from a controlled within-record perspective. Ali et al. [20] examined CHB-MIT at a broader cross-subject scale and emphasized the influence of segment definition, channel choice, class balance, and patient-wise partitioning. Brookshire et al. [21] demonstrated how segment-level splitting can leak subject- or session-specific information across EEG training and test sets. The reduction in SVM MCC from 0.911 under random-window validation to 0.784 under block-preserving validation provides a direct record-level example of that mechanism. In contrast to the DWT-CNN-LSTM detector of Kashefi Amiri et al. [23], the present study deliberately uses probabilistic, kernel, and explicit fuzzy models to isolate methodological effects rather than maximize headline accuracy.

The principal contribution of this study is the joint experimental evaluation of annotation fidelity, temporal grouping, preprocessing placement, and fuzzy-model specification within a single controlled pipeline. Examining these factors together reveals that apparent gains attributed to a classifier may partly originate from data construction or validation choices. This integrated design provides a reproducible basis for distinguishing algorithmic improvement from methodological optimism.

Several limitations remain important. Only one single-channel record was analyzed, and the interictal intervals were selected from the same recording. Differences between the interictal and ictal segments may therefore reflect artifacts, vigilance, or amplitude context in addition to seizure physiology. The record was not evaluated continuously, so event sensitivity, detection delay, calibration, and false alarms per hour were not established. The preprocessing comparison was not nested, and the pediatric CHB-MIT population limits broader clinical interpretation. Future work should repeat the complete pipeline across multiple subjects, preserve patient identity in outer folds, tune preprocessing and hyperparameters in inner folds, and report event-level outcomes [2,15]. The buffered stress test additionally showed that performance is unstable when temporal separation is made more conservative, and part of that decline is inseparable from the accompanying reduction in training data in this compact design. The preprocessing comparison was not nested and therefore remains descriptive rather than a basis for selecting a superior configuration. These constraints reinforce that the present work is a methodological single-record case study, not a seizure detector validated for clinical deployment.

Conclusion

Annotation fidelity, temporal grouping, and fuzzy-model specification materially affected wavelet-based EEG seizure-classification results in the analyzed

record. Corrected fixed-window features, fold-wise scaling, temporal block preservation, and an explicit first-order TSK model produced a coherent leakage-aware pipeline. Random-window evaluation was more optimistic than block-preserving evaluation, while wavelet family and window design changed the balance between sensitivity and overall classification performance. The study therefore demonstrates that data construction and validation design must be treated as core experimental factors rather than implementation details. Multi-record and patient-independent testing is required before the observed performance can be generalized clinically. The temporal-buffer stress test further showed that estimates can deteriorate markedly when local temporal proximity is removed. Accordingly, the reported values should be interpreted only as within-record methodological evidence.

Acknowledgment

The authors acknowledge the investigators and contributors of the CHB-MIT Scalp EEG Database and PhysioNet for providing de-identified EEG recordings for research.

References

1. Shoeb AH, Guttag JV. Application of machine learning to epileptic seizure detection. In: Proceedings of the 27th International Conference on Machine Learning; 2010. p. 975-82.
2. Boonyakitanont P, Lek-uthai A, Chomtho K, Songsiri J. A review of feature extraction and performance evaluation in epileptic seizure detection using EEG. Biomed Signal Process Control. 2020;57:101702.
3. Adeli H, Zhou Z, Dadmehr N. Analysis of EEG records in an epileptic patient using wavelet transform. J Neurosci Methods. 2003;123(1):69-87.
4. Guler I, Ubeyli ED. Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficients. J Neurosci Methods. 2005;148(2):113-21.
5. Subasi A. Application of adaptive neuro-fuzzy inference system for epileptic seizure detection using wavelet feature extraction. Comput Biol Med. 2007;37(2):227-44.
6. Gandhi T, Panigrahi BK, Anand S. A comparative study of wavelet families for EEG signal classification. Neurocomputing. 2011;74(17):3051-7.
7. Xie S, Krishnan S. Wavelet-based sparse functional linear model with applications to EEG seizure detection and epilepsy diagnosis. Med Biol Eng Comput. 2013;51(1-2):49-60.
8. Raghu S, Sriraam N. Classification of focal and non-focal EEG signals using neighborhood component analysis and machine learning algorithms. Expert Syst Appl. 2018;113:18-32.
9. Shoeibi A, Ghassemi N, Khodatars M, Moridian P, Alizadehsani R, Zare A, et al. Detection of epileptic seizures on EEG signals using ANFIS classifier, autoencoders and fuzzy entropies. Biomed Signal Process Control. 2022;73:103417.
10. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res. 2010;11:2079-107.
11. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91.
12. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1-30.
13. Yang C, Deng Z, Choi KS, Wang S. Takagi-Sugeno-Kang transfer learning fuzzy logic system for adaptive recognition of epileptic electroencephalogram signals. IEEE Trans Fuzzy Syst. 2016;24(5):1079-94.
14. Jiang Y, Deng Z, Chung FLK, Wang G, Qian P, Choi KS, et al. Recognition of epileptic EEG signals using a novel multiview TSK fuzzy system. IEEE Trans Fuzzy Syst. 2017;25(1):3-20.
15. Lee K, Jeong H, Kim S, Yang D, Kang HC, Choi E. Real-time seizure detection using EEG: a comprehensive comparison of recent approaches under a realistic setting. Proc Mach Learn Res. 2022;174:311-37.
16. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource. Circulation. 2000;101(23):e215-20.
17. Guttag J. CHB-MIT Scalp EEG Database (version 1.0.0). PhysioNet. 2010. doi:10.13026/C2K01R.
18. Sharma R, Pachori RB, Acharya UR. An integrated index for identification of focal electroencephalogram signals using discrete wavelet transform and entropy measures. Entropy. 2015;17(8):5218-40.
19. Li Y, Wang X, Luo M, Li K, Yang X, Guo Q. Epileptic seizure classification of EEGs using time-frequency analysis based multiscale radial basis functions. IEEE J Biomed Health Inform. 2018;22(2):386-97.
20. Ali E, Angelova M, Karmakar C. Epileptic seizure detection using CHB-MIT dataset: the overlooked perspectives. R Soc Open Sci. 2024;11(5):230601. doi:10.1098/rsos.230601.
21. Brookshire G, Kasper J, Blauch NM, Wu YC, Glatt R, Merrill DA, et al. Data leakage in deep learning studies of translational EEG. Front Neurosci. 2024;18:1373515. doi:10.3389/fnins.2024.1373515.
22. Zhang X, Zhang X, Huang Q, Chen F. A review of epilepsy detection and prediction methods based on EEG signal processing and deep learning. Front Neurosci. 2024;18:1468967. doi:10.3389/fnins.2024.1468967.
23. Kashefi Amiri H, Zarei M, Daliri MR. Epileptic seizure detection from electroencephalogram signals based on 1D CNN-LSTM deep learning model using discrete wavelet transform. Sci Rep. 2025;15:32820. doi:10.1038/s41598-025-18479-9.