I. J. Engineering and Manufacturing, **2026, 5, 179-204**
Published Online on October 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2026.05.10

Leakage-Controlled Recording-Level Evaluation of Machine Learning, Deep Learning, and Convolutional-Kernel Models for EEG-Based Epileptic Seizure Detection

Sanagavarapu Sunitha*
School of Science and Computer Studies, CMR University, Bengaluru, Karnataka, India
E-mail: sanagavarapu.sunitha@cmr.edu.in
ORCID iD: https://orcid.org/0000-0003-1167-2112
*Corresponding Author

Umadevi Ramamoorthy
School of Science and Computer Studies, CMR University, Bengaluru, Karnataka, India
E-mail: umadevi.r@cmr.edu.in
ORCID iD: https://orcid.org/0000-0002-6234-3573

Received: July 7, 2026; Revised: July 16, 2026; Accepted: August 17, 2026; Published: October 8, 2026

Abstract: Electroencephalogram (EEG)-based seizure detection is a challenging time-series classification problem because EEG signals are nonlinear, non-stationary, and temporally heterogeneous. This study presents a unified comparison of classical machine-learning, neural and deep-learning, convolutional-kernel time-series, and probability-level ensemble models using the Epileptic Seizure Recognition benchmark dataset. The 11,500 EEG segments were reconstructed into 500 recording groups, and recording-grouped nested cross-validation was applied to prevent segments from the same recording from crossing training and test partitions. Depending on the modelling paradigm, segments were represented either by 67 handcrafted time-, frequency-, and nonlinear-domain features or by ordered 178-sample EEG sequences. Recording-level average precision was used as the primary evaluation metric, supported by ROC-AUC, threshold-dependent measures, Brier score, bootstrap confidence intervals, ablation analysis, explainability, and computational-efficiency benchmarking. RBF-SVM achieved the highest average precision of 0.997985 and ROC-AUC of 0.999475. MultiRocket produced the lowest family-representative Brier score of 0.006148, while MLP achieved the lowest standalone inference latency after feature extraction. The selected meta-ensemble produced strong threshold-dependent performance but did not improve overall precision–recall ranking. Paired bootstrap analysis found no statistically significant differences among the family representatives after multiple-comparison correction. These findings demonstrate that handcrafted-feature and time-series models can both achieve strong recording-level performance when evaluated under a consistent, leakage-controlled protocol. Because verified patient identifiers were unavailable, the results should not be interpreted as evidence of patient-independent clinical generalization.

Index Terms: EEG seizure detection, recording-grouped cross-validation, leakage-controlled evaluation, machine learning, deep learning, convolutional-kernel time-series classification, computational efficiency

1. Introduction

Epilepsy is a chronic neurological disorder characterized by recurrent, unpredictable seizures that can substantially impair safety, independence, and quality of life. Accurate and timely seizure detection is therefore important for diagnosis, monitoring, and treatment assessment. Electroencephalography (EEG) remains the principal non-invasive modality for observing brain electrical activity because of its high temporal resolution and sensitivity to abnormal neural discharges. Artificial-intelligence methods have increasingly been applied to automate EEG interpretation and distinguish seizure from non-seizure activity. Ensemble learning combined with deep architectures has shown potential for improving seizure-detection performance [1], while scalable time-series classifiers have demonstrated the value of modelling temporal structure efficiently in electrophysiological signals [2].

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

EEG classification remains difficult because epileptic activity is nonlinear, non-stationary, and temporally heterogeneous. Lytton [3] described epilepsy as a dynamic-systems problem, supporting the use of models capable of capturing complex temporal dependencies beyond simple statistical assumptions. Deep-learning methods for time-series anomaly detection further demonstrate the ability of learned representations to identify subtle and transient deviations in sequential data [4]. However, greater model complexity does not automatically ensure better discrimination, calibration, interpretability, or computational efficiency. Broader healthcare-AI research likewise emphasizes that diagnostic models require rigorous, transparent, and clinically cautious evaluation rather than accuracy-only reporting [5].

Existing EEG studies are difficult to compare because they often differ in partitioning granularity, preprocessing, feature engineering, model input representation, probability aggregation, threshold selection, and reported metrics. Segment-level random splitting can also place windows from the same original recording in both training and test sets, producing optimistic estimates. Moreover, many studies evaluate only one methodological family, making it unclear whether improvements arise from the classifier, signal representation, or evaluation protocol. The Epileptic Seizure Recognition dataset [6] provides a common benchmark for comparative analysis, but its segmented structure requires reconstruction and recording-grouped validation to prevent intra-record leakage.

This study addresses these limitations through a unified, leakage-controlled framework for comparing classical machine-learning, neural and deep-learning, convolutional-kernel time-series, and probability-level ensemble models. The processed segments are reconstructed into their original recording groups, and model development is conducted using recording-grouped nested cross-validation. Depending on the modelling paradigm, EEG segments are represented either by 67 handcrafted time-, frequency-, and nonlinear-domain features or by ordered 178-sample raw sequences. Models are compared using recording-level average precision as the primary metric, supported by ROC-AUC, threshold-dependent measures, probability calibration, bootstrap confidence intervals, paired statistical tests, ablation analysis, explainability, and computational-efficiency benchmarking.

The principal contribution is therefore not a new classifier in isolation, but a standardized comparative protocol that separates the effects of representation, model family, aggregation, and operating threshold. By jointly evaluating predictive performance, uncertainty, probability quality, interpretability, and efficiency–accuracy trade-offs, the study provides a rigorous basis for identifying practically suitable EEG seizure-detection systems. Because verified patient identifiers are unavailable, the findings are interpreted as recording-disjoint benchmark results rather than evidence of patient-independent clinical generalization.

2. Literature Review

Computational analysis of neurological signals has advanced substantially over the past two decades, progressing from physiological interpretation towards automated EEG classification. Magnuson et al. [7] examined abnormal neural activity associated with traumatic brain injury and highlighted the complex and nonlinear nature of disturbed brain dynamics. Although their study did not focus on epilepsy, it provides methodological support for analytical approaches capable of detecting subtle temporal variations in neurological signals. More directly, Kode et al. [8] evaluated machine-learning and deep-learning techniques for EEG seizure classification and demonstrated that model selection, signal representation, and preprocessing strongly influence predictive performance. Mohammadi Foumani et al. [9] extended this perspective through a broad review of deep-learning methods for time-series classification and regression, showing how convolutional and recurrent architectures can learn hierarchical representations directly from sequential data.

Specialised architectures have also been investigated for biomedical time-series analysis. Liu et al. [10] proposed a three-dimensional convolutional architecture for medical classification. Although their application concerned lung cancer rather than EEG seizure detection, the study illustrates the potential of jointly modelling spatial and temporal information. In epilepsy-focused research, Jahan et al. [11] reviewed AI-based seizure detection and prediction within Internet of Healthcare Things systems and identified limitations related to generalisation, interpretability, and practical implementation. This distinction between benchmark accuracy and operational reliability is particularly important for clinical and wearable monitoring systems.

Semi-supervised and structure-aware approaches have been proposed to reduce dependence on extensively labelled datasets. Chen et al. [12] introduced trend- and order-based semi-supervised time-series classifiers designed to exploit temporal organisation under limited supervision. Earlier, Tzallas et al. [13] reviewed automated seizure-detection systems based on signal processing, handcrafted feature extraction, and conventional classification, documenting the transition from rule-based methods towards data-driven learning.

More recent reviews have examined the relative value of deep and classical learning approaches. Shoeibi et al. [14] reported that convolutional and recurrent models can achieve strong seizure-detection performance, while also identifying increased computational demand, architecture sensitivity, and interpretability limitations. Siddiqui et al. [15] similarly concluded that conventional machine-learning classifiers remain competitive when informative EEG features are available or computational resources are constrained. Vidyaratne and IftekharUddin [16] further emphasised the importance of latency and computational cost for real-time seizure detection, whereas Tran et al. [17] identified inconsistent preprocessing, partitioning, and evaluation practices as major barriers to meaningful comparison across

studies. Yu et al. [18] additionally demonstrated the growing relevance of AI-assisted wearable systems, reinforcing the need to evaluate sensitivity, reliability, and computational feasibility together.

Earlier studies provide important context for the development of current seizure-detection methods. Paul [19] reviewed seizure-detection techniques based on different biomedical signals, while Sadati et al. [20] applied neural-fuzzy networks to epileptic seizure detection. Guo et al. [21] combined line-length features with artificial neural networks, illustrating the continued value of handcrafted signal descriptors. Many subsequent studies have relied on the EEG benchmark introduced by Andrzejak et al. [22], which remains widely used for evaluating seizure-classification methods.

Despite these developments, existing studies frequently differ in signal normalisation, feature engineering, partitioning granularity, aggregation rules, threshold selection, and reported metrics. Segment-level partitioning may produce optimistic estimates when segments originating from the same recording occur in both training and test sets. Furthermore, many studies evaluate only one model family and omit calibration, statistical uncertainty, ablation analysis, interpretability, or computational requirements. The literature therefore supports the need for a unified, recording-grouped comparison of classical machine learning, deep learning, convolutional-kernel time-series models, and ensemble methods. The present study addresses this gap by standardising partitioning, preprocessing, probability aggregation, evaluation metrics, statistical analysis, interpretability assessment, and efficiency benchmarking within a common experimental framework.

3. Methodology

A unified leakage-safe framework was developed to compare classical machine-learning, deep-learning, convolutional-kernel time-series, and ensemble models for binary epileptic-seizure detection. The framework comprised reconstruction of the source-recording structure, recording-grouped nested cross-validation, fold-specific preprocessing, handcrafted and raw-signal representations, model and threshold selection, recording-level probability aggregation, statistical comparison, ablation analysis, explainability assessment, and computational-efficiency evaluation.

3.1. Dataset Description and Recording Reconstruction

The publicly available Epileptic Seizure Recognition dataset was used in this study [6]. It is a preprocessed version of the EEG benchmark originally introduced by Andrzejak et al. and contains 500 single-channel EEG recordings distributed equally across five physiological categories. Each original recording consists of 4,097 EEG amplitude measurements acquired over approximately 23.6 s.

In the processed CSV file, each recording is represented by 23 segments, with each segment containing 178 temporally ordered EEG values ($X_1 - X_{178}$). The complete dataset therefore contains 11,500 segment-level observations. The 178 values were treated as an ordered EEG sequence rather than as independent tabular variables.

The original five classes were converted into a binary seizure-detection problem. Class 1 represented seizure activity, whereas Classes 2–5 were combined into the non-seizure category:

$$y_i^{(b)} = \begin{cases} 1, & y_i = 1, \\ 0, & y_i \in \{2, 3, 4, 5\}, \end{cases} \tag{1}$$

where y_i is the original five-class label and $y_i^{(b)}$ is the corresponding binary label.

The resulting dataset comprised 100 seizure and 400 non-seizure recordings, corresponding to 2,300 seizure and 9,200 non-seizure segments. Because verified patient identifiers were unavailable, the reconstructed groups were treated as recordings rather than independent patients.

Before model development, the 11,500 rows were assigned to 500 recording groups, with 23 ordered segments per recording. Recording identifiers and segment-position indices were retained only for grouped partitioning, probability aggregation, and error analysis. Integrity checks confirmed the expected recording counts, consistent labels within each recording, absence of duplicate recording–segment combinations, and absence of missing or infinite signal values.

3.2. Data Preprocessing and Leakage-Safe Partitioning

3.2.1. Initial Data Processing

The non-predictive identifier column was removed from the feature matrices. The 178 EEG values were retained in their original sequence, and the five-class response variable was converted to the binary seizure/non-seizure label defined in Eq. (1).

No additional band-pass filtering, notch filtering, artefact rejection, or signal decomposition was applied because the publicly released data had already been preprocessed and segmented. The resulting findings therefore represent performance on the processed benchmark dataset rather than on unfiltered clinical EEG recordings.

Two data representations were generated from every segment. The first consisted of 67 handcrafted EEG features. The second retained the raw sequence of 178 amplitude values. All feature extraction procedures were fixed before model evaluation and were applied consistently across the cross-validation folds.

3.2.2. Grouped Nested Cross-Validation

A five-fold grouped nested cross-validation strategy was adopted to prevent segments derived from the same recording from appearing in both training and evaluation data. The recording identifier was used as the grouping variable in every partitioning operation.

In the outer loop, the 500 recordings were divided into five stratified grouped folds. Each outer-test fold contained approximately 100 recordings, including approximately 20 seizure and 80 non-seizure recordings. All 23 segments belonging to a given recording were assigned to the same outer fold. Consequently, a recording could contribute data either to the outer-training partition or to the outer-test partition, but never to both.

For each outer iteration, the remaining approximately 400 recordings were divided into five grouped inner folds. The inner folds were used for:

- hyperparameter and architecture selection;
- preprocessing selection;
- deep-model epoch determination;
- recording-level aggregation selection;
- decision-threshold selection;
- ensemble-weight estimation; and
- probability-calibration selection.

The outer-test partition was not used for any of these decisions. After the optimal configuration had been identified from the inner folds, the selected model was refitted using the complete outer-training partition and evaluated once on the corresponding outer-test recordings. Pooling the five outer-test folds produced one unbiased out-of-fold recording-level prediction for each of the 500 recordings.

3.2.3. Fold-Specific Scaling and Normalisation

All scaling and normalisation parameters were estimated exclusively from the relevant training partition. For a feature value x_{ij}, standardisation was performed as

$$z_{ij} = \frac{x_{ij} - \mu_j^{(\text{train})}}{\sigma_j^{(\text{train})}}, \tag{2}$$

where $\mu_j^{(\text{train})}$ and $\sigma_j^{(\text{train})}$ are the mean and standard deviation of feature j, respectively, calculated only from the current training fold. The resulting transformation was applied without re-estimation to the corresponding validation or test observations.

The handcrafted-feature representation was standardised before fitting Logistic Regression, RBF-SVM, k-NN, Gaussian Naive Bayes, and MLP. Tree-based models were evaluated within their fold-specific feature pipelines. For raw-sequence models, the signal mean and scale were calculated from the training partition and applied to the validation or test sequences. In the ROCKET family, the resulting kernel-transformed features were additionally standardised before fitting the linear probability classifier.

3.3. Feature Representation

3.3.1. Handcrafted EEG Features

A 67-dimensional handcrafted feature vector was extracted independently from each EEG segment, comprising 28 time-domain, 30 frequency-domain, and 9 nonlinear features.

The time-domain features characterised signal magnitude, distribution, variability, waveform shape, and temporal fluctuation. For an EEG segment $x = \{x_1, x_2, \ldots, x_N\}$, where $N = 178$, the sample mean and variance were calculated as

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n\,,\; s^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2. \tag{3}$$

Signal energy was calculated as

$$E = \sum_{n=1}^{N} x_n^2. \tag{4}$$

The remaining time-domain features included root-mean-square amplitude, mean absolute value, median, quartiles, interquartile range, median absolute deviation, skewness, excess kurtosis, zero-crossing measures, mean absolute difference, line length, and related waveform-complexity measures.

The frequency-domain features described the spectral composition of each segment and included absolute and relative band powers, dominant-frequency measures, spectral centroid, spectral spread, spectral entropy, spectral slope, and selected band-power ratios. The relative power of frequency band B was calculated as

$$RP_B = \frac{\sum_{f \in B} P(f)}{\sum_{f \in F} P(f)}, \tag{5}$$

where $P(f)$ denotes the estimated power spectral density, B represents the frequency band of interest, and F denotes the complete analysed frequency range.

The nonlinear features characterised signal irregularity, predictability, long-range dependence, and fractal structure. They included approximate entropy, sample entropy, permutation-based complexity, Hurst-type scaling, and fractal-dimension measures. All feature definitions and calculation parameters were fixed before model evaluation and applied consistently across the cross-validation folds.

3.3.2. Raw EEG Representation

The raw representation retained the 178 temporally ordered EEG values from each segment. Raw-input models received an input tensor of shape 178×1, corresponding to 178 time positions and one signal channel. This representation allowed convolutional, recurrent, and convolutional-kernel models to learn discriminative temporal patterns without prior handcrafted summarisation.

The MLP used the 67-dimensional handcrafted representation. CNN-1D, SimpleRNN, LSTM, TCN, InceptionTime, EEGNet-1D, ROCKET, MiniRocket, and MultiRocket used the normalised raw EEG sequence. Table 1 presents the complete model–feature mapping used in the comparative framework.

Table 1. Model–feature mapping

Model family	Evaluated models	Input representation	Input dimension
Classical machine learning	Logistic Regression, RBF-SVM, k-NN, Random Forest, Extra Trees, Gradient Boosting, Gaussian Naive Bayes	Handcrafted EEG features	67
Feature-based neural learning	MLP	Handcrafted EEG features	67
Raw-sequence deep learning	CNN-1D, SimpleRNN, LSTM, TCN, InceptionTime, EEGNet-1D	Normalised raw EEG sequence	178×1
Convolutional-kernel time series	ROCKET, MiniRocket, MultiRocket	Normalised raw EEG sequence	178×1
Meta-ensemble	Recording-level probabilities from 17 base models	Probability vector	Up to 17

3.4. Model Implementation

A total of 18 predictive systems were evaluated: seven classical machine-learning models, seven neural and deep-learning models, three convolutional-kernel models, and one probability-level ensemble.

3.4.1. Classical Machine-Learning Models

The classical model family consisted of Logistic Regression, RBF-SVM, k-Nearest Neighbours, Random Forest, Extra Trees, Gradient Boosting, and Gaussian Naive Bayes.

Logistic Regression provided a regularised linear probabilistic baseline. RBF-SVM modelled nonlinear class boundaries through the radial basis function kernel. k-NN classified observations according to their neighbouring feature vectors. Random Forest and Extra Trees represented bagging-based tree ensembles with different levels of split randomisation. Gradient Boosting constructed an additive sequence of decision trees, while Gaussian Naive Bayes provided a distribution-based probabilistic baseline.

Model-specific candidate configurations were evaluated in the grouped inner folds. The investigated settings included regularisation strength for Logistic Regression; regularisation and kernel-width parameters for RBF-SVM; neighbour count, distance metric, and weighting strategy for k-NN; estimator count, maximum tree depth, feature subsampling, and minimum node size for Random Forest and Extra Trees; estimator count, learning rate, and tree complexity for Gradient Boosting; and variance smoothing for Gaussian Naive Bayes.

The configuration with the strongest mean inner recording-level average precision was selected within each outer-training partition. ROC-AUC and calibration performance were used to distinguish configurations with closely matched average precision.

3.4.2. Neural and Deep-Learning Models

Seven neural architectures were implemented in PyTorch: MLP, CNN-1D, SimpleRNN, LSTM, TCN, InceptionTime, and EEGNet-1D. The MLP processed the 67 standardised handcrafted features using fully connected layers and dropout regularisation. The remaining architectures used the normalised 178×1EEG sequence. CNN-1D learned local temporal patterns through one-dimensional convolutions; SimpleRNN and LSTM modelled sequential dependencies using recurrent hidden states and gated memory, respectively. TCN employed stacked dilated convolutions, InceptionTime used parallel convolutional branches with different receptive fields, and EEGNet-1D applied compact temporal and separable convolutional operations based on the EEGNet architecture [23].

Three predefined candidate configurations were evaluated for each architecture. These configurations varied hidden-layer width, convolutional channels, recurrent units, kernel or receptive-field settings, dropout, learning rate, and weight decay. Model configuration was selected independently within each outer-training partition using mean recording-level average precision across the grouped inner folds.

Binary cross-entropy with fold-specific positive-class weighting was used to address class imbalance. The class weight was calculated only from the seizure and non-seizure samples in the corresponding training partition. AdamW was used for optimisation, gradient-norm clipping was applied to stabilise training, and all computations used float32 precision.

Within each inner fold, the epoch yielding the highest validation recording-level average precision was retained. The median of the five inner-fold best epochs determined the final training duration for the selected configuration in the corresponding outer fold. Final training durations ranged from 8 to 26 epochs. In total, the nested deep-learning evaluation involved 525 inner training runs, obtained from seven architectures, three candidate configurations, five outer folds, and five inner folds.

3.4.3. Convolutional-Kernel Time-Series Models

ROCKET [24], MiniRocket [25], and MultiRocket [26] were evaluated as convolutional-kernel time-series classifiers. Each method transformed the normalised 178×1 EEG sequence into a high-dimensional feature representation using convolutional kernels with different dilations, biases, and response statistics. ROCKET used randomly parameterised kernels, MiniRocket employed a more deterministic and computationally efficient transformation, and MultiRocket generated multiple response summaries from complementary signal representations.

Within each outer fold, signal normalisation and the convolutional-kernel transformation were fitted using only the outer-training data. The transformed features were subsequently standardised, and regularised Logistic Regression was used to generate seizure probabilities.

Candidate configurations varied the kernel budget, dilation settings, number of transformed features, signal normalisation, Logistic Regression regularisation, class weighting, solver, convergence tolerance, and maximum iterations. The evaluated MultiRocket configurations used either 504 or 2,520 kernels, producing 4,032 or 20,160 transformed features, respectively. Configuration selection was performed using recording-level average precision obtained from the grouped inner folds.

3.4.4. Leakage-Safe Meta-Ensemble

The meta-ensemble combined recording-level seizure probabilities generated by the 17 base models. It was constructed independently within each outer-training partition using only inner out-of-fold predictions. Consequently, the corresponding outer-test labels and probabilities were not used for base-model selection, weight estimation, calibration, aggregation selection, or decision-threshold optimisation.

For recording $r$in outer fold f, the ensemble probability was calculated as

$$p_{r,f}^{(\mathrm{ens})} = \sum_{m=1}^{17} w_{m,f}\, p_{r,m,f},\ w_{m,f} \geq 0,\ \sum_{m=1}^{17} w_{m,f} = 1, \tag{6}$$

where $p_{r,m,f}$ denotes the seizure probability produced by base model m, and $w_{m,f}$ denotes its fold-specific ensemble weight. A model was considered active when $w_{m,f} > 0$; therefore, zero-valued weights allowed uninformative probability sources to be excluded.

Equal-weight averaging and optimised non-negative weighting were evaluated using the inner out-of-fold recording predictions. Isotonic regression was also considered for probability calibration and was retained only when it improved the predefined inner-validation probabilistic criterion. The numbers of active models in outer folds F1–F5 were 3, 4, 2, 17, and 5, respectively. Calibration was not retained in F1, F3, F4, or F5, whereas isotonic calibration was selected in F2. The active models, probability weights, calibration method, recording-level aggregation rule, and classification threshold were selected separately within each outer fold. Table 2 consolidates the exact fold-specific configurations retained in the final nested evaluation.

Table 2. Model implementation and configuration controls

Model	Principal implementation	Exact selected configuration in F1–F5
Logistic Regression	Regularised linear probability model	(C=) **10 / 0.1 / 1 / 0.1 / 10**
RBF-SVM	Radial-basis kernel classifier	(C=5) throughout; (\gamma=) scale / scale / scale / 0.01 / scale; probability estimation enabled
k-NN	Distance-based classifier	Neighbours = 3 / 3 / 5 / 3 / 3; weights = distance / distance / distance / uniform / uniform
Random Forest	Bagged decision-tree ensemble	Maximum depth = 24 / None / None / None / None; minimum leaf size = 1 / 1 / 1 / 1 / 2
Extra Trees	Randomised decision-tree ensemble	Maximum depth = 24 / None / None / None / None; minimum leaf size = 1 / 1 / 1 / 1 / 1
Gradient Boosting	Sequential boosted trees	((n_{\mathrm{estimators}},\eta,\mathrm{depth})=) (200, 0.10, 2) / (100, 0.05, 2) / (200, 0.05, 3) / (200, 0.05, 2) / (200, 0.10, 3)
Gaussian Naive Bayes	Gaussian probabilistic classifier	Variance smoothing = (1\times10^{-11}) throughout
MLP	Fully connected neural network	Candidate = 002 / 001 / 001 / 001 / 001; epochs = 8 / 8 / 8 / 8 / 8; trainable parameters = 59,521 / 17,409 / 17,409 / 17,409 / 17,409
CNN-1D	One-dimensional convolutional network	Candidate = 002 / 001 / 001 / 001 / 001; epochs = 11 / 8 / 12 / 11 / 8
SimpleRNN	Recurrent hidden-state network	Candidate = 002 / 001 / 003 / 003 / 003; epochs = 8 / 11 / 14 / 11 / 10
LSTM	Gated recurrent network	Candidate = 003 / 001 / 003 / 003 / 003; epochs = 11 / 21 / 12 / 14 / 11
TCN	Dilated temporal convolutional network	Candidate = 001 / 001 / 002 / 001 / 002; epochs = 13 / 9 / 8 / 13 / 9
InceptionTime	Parallel multi-scale convolutional network	Candidate = 002 / 001 / 003 / 001 / 002; epochs = 11 / 10 / 11 / 9 / 10
EEGNet-1D	Compact temporal and separable CNN	Candidate = 003 / 001 / 001 / 003 / 003; epochs = 11 / 12 / 20 / 26 / 13
ROCKET	Random convolutional transformation	Transformed features = 10,000 / 4,000 / 10,000 / 4,000 / 4,000; equivalent kernels = 5,000 / 2,000 / 5,000 / 2,000 / 2,000
MiniRocket	Deterministic convolutional transformation	Transformed features = 9,996 / 9,996 / 4,032 / 4,032 / 19,992; maximum dilations per kernel = 32
MultiRocket	Multi-representation convolutional transformation	Kernels = 504 / 2,520 / 2,520 / 504 / 2,520; transformed features = 4,032 / 20,160 / 20,160 / 4,032 / 20,160; four features per kernel; maximum dilations = 32
Meta-ensemble	Weighted recording-level probability combination	Active models = 3 / 4 / 2 / 17 / 5; calibration = none / isotonic / none / none / none; non-negative weights and thresholds selected from inner out-of-fold predictions

Note: F1–F5 represent outer folds 1–5, while "None" denotes unrestricted tree depth. Classical-model hyperparameters were selected using grouped inner-fold recording-level average precision. Neural configurations were chosen from three predefined candidates, with final training epochs determined by the median inner-fold optimum. Convolutional-kernel models used standardised transformed features and balanced L2-regularised Logistic Regression with $C = 1$, the liblinear solver, 3,000 maximum iterations, and a tolerance of 10^{-4}.

3.5. Training and Evaluation Protocol

3.5.1. Recording-Level Probability Aggregation

Each base model initially generated one seizure probability for each of the 23 segments belonging to a reconstructed recording. These segment probabilities were aggregated into a single recording-level probability before final evaluation.

Eight aggregation strategies were examined:

1. arithmetic mean;
2. median;
3. maximum;
4. mean of the three largest probabilities;

5. mean of the five largest probabilities;
6. upper-quartile mean;
7. logit mean; and
8. proportion of positively classified segments.

The aggregation method was selected using inner out-of-fold recording predictions only. The selected rule was then fixed before evaluation on the corresponding outer-test fold. The final family-level analysis used mean aggregation for RBF-SVM, logit-mean aggregation for MLP, and median aggregation for MultiRocket.

3.5.2. Decision-Threshold Selection

A recording was classified as seizure when its aggregated probability was greater than or equal to the selected threshold τ:

$$\hat{y}_r = \begin{cases} 1, & p_r \geq \tau, \\ 0, & p_r < \tau, \end{cases} \tag{7}$$

where p_r is the aggregated recording-level probability.

Candidate thresholds were evaluated using the inner out-of-fold predictions. The threshold maximising recording-level Matthews correlation coefficient was selected. Balanced accuracy, sensitivity, and specificity were used as secondary criteria where necessary. No threshold was selected or adjusted using outer-test outcomes.

3.5.3. Performance Measures

The reconstructed EEG recording was treated as the primary unit of evaluation. Segment-level predictions were retained for recording-level aggregation, attribution analysis, and diagnostic assessment but were not considered independent observations in the principal cross-family comparison.

Performance was evaluated using average precision, ROC-AUC, accuracy, balanced accuracy, sensitivity, specificity, precision, F1-score, Matthews correlation coefficient, Brier score, and false-positive and false-negative counts. Recording-level calibration curves were additionally examined to assess the agreement between predicted probabilities and observed outcomes.

Average precision was used as the primary ranking metric because seizure recordings constituted 20% of the recording-level dataset. ROC-AUC assessed discrimination across classification thresholds, MCC summarized all four components of the confusion matrix, and the Brier score quantified the accuracy of the predicted probabilities.

3.5.4. Confidence Intervals and Paired Statistical Comparisons

Uncertainty was estimated using 2,000 stratified bootstrap repetitions of the pooled outer-fold recording predictions. Seizure and non-seizure recordings were resampled separately with replacement, preserving the original recording-level class composition. Percentile-based 95% confidence intervals were calculated for the principal discrimination, classification, and calibration measures.

Paired bootstrap comparisons were performed using identical resampled recording indices for both models in each comparison. Differences in average precision, ROC-AUC, MCC, and Brier score were calculated for every bootstrap repetition. Two-sided empirical p-values were derived from the resulting difference distributions, and Holm adjustment was applied to control the family-wise error rate across multiple comparisons.

Exact McNemar tests were additionally performed for selected model pairs using their recording-level binary decisions. The test considered recordings misclassified by one model but correctly classified by the other and therefore assessed whether the paired error patterns differed beyond random variation.

3.5.5. Ablation Analyses

Structured ablation analyses were conducted to examine the effects of feature representation, preprocessing, recording-level aggregation, and convolutional-kernel configuration. For the reference RBF-SVM, the complete 67-feature representation was compared with time-domain-only, frequency-domain-only, nonlinear-only, and leave-one-domain-out feature sets. Standard scaling, robust scaling, and no scaling were also evaluated.

Recording-level decision formation was assessed by comparing eight probability-aggregation strategies. Kernel-model ablations examined the transformation type, kernel budget, generated-feature dimensionality, Logistic Regression regularisation, feature selection, and class weighting. Model and preprocessing choices were determined using training or grouped inner-validation data, while pooled outer-test predictions were used only for final reporting.

3.5.6. Explainability Analysis

The reference classical model was interpreted through recording-block permutation importance. Rather than permuting individual segment rows independently, all segments belonging to one recording were treated as a block, thereby preserving the hierarchical structure of the data. Each of the 67 handcrafted features was permuted ten times

within each outer fold. The resulting changes in average precision, ROC-AUC, MCC, and Brier score were used to quantify feature importance.

The seven neural models were examined using input-attribution analysis. Twenty recordings were selected for each model and outer fold, producing 700 model–recording attribution evaluations. Both correctly classified and erroneous cases were represented.

Attribution faithfulness was assessed by masking the most highly attributed 10%, 20%, and 30% of input positions and measuring the resulting reduction in prediction confidence. Random masking at the same proportions was used as a reference. Stability of the global attribution profiles across outer folds was evaluated using Spearman rank correlation.

3.5.7. Computational-Efficiency Evaluation

Computational efficiency was evaluated for the selected classical, neural, convolutional-kernel, and ensemble representatives using their saved outer-fold artifacts. The benchmark measured model-loading time, full-fold inference latency, throughput, Python peak memory, resident-memory change, serialized artifact size, and model-specific complexity.

Each saved artifact was applied to its complete outer-test fold, and the reproduced probabilities were compared with the authoritative stored predictions to verify numerical consistency. Loading time was measured separately from inference latency. Complexity was represented by the trainable-parameter count for MLP, support-vector count for RBF-SVM, kernel and transformed-feature counts for MultiRocket, and number of non-zero weights for the ensemble.

The reported ensemble latency included only probability weighting and optional calibration after the base-model probabilities had been generated. It therefore represented incremental combination overhead rather than complete end-to-end ensemble inference and was not directly comparable with the full latency of an individual model.

3.6. Experimental Environment and Reproducibility

The experiments were conducted in a Windows-based Python environment. NumPy and pandas were used for numerical and tabular processing, scikit-learn for preprocessing, classical machine-learning models, calibration, and statistical analysis, PyTorch for neural and deep-learning models, and aeon for ROCKET-family transformations. Deep-learning computations were performed using an NVIDIA GeForce RTX 3060 GPU with CUDA acceleration and float32 precision. The maximum GPU-memory usage recorded during nested training was approximately 2.64 GB.

Reproducibility was supported through fixed stage-specific random seeds, predefined recording-group assignments, stored inner- and outer-fold partitions, fold-specific preprocessing parameters, selected hyperparameters, training-epoch counts, architecture configurations, trained model artifacts, recording-level aggregation rules, classification thresholds, ensemble weights, calibration objects, prediction files, and execution manifests.

Each computational stage generated a checkpoint containing its configuration, execution status, input and output information, and run signature. SHA-256 checksums were generated for key artifacts to verify file integrity. Prediction-reproduction audits were also performed to confirm that the saved model artifacts regenerated the corresponding outer-fold probabilities within the predefined numerical tolerance.

4. Results and Discussion

The experimental results were evaluated at the reconstructed recording level using the pooled out-of-fold predictions obtained from the five outer cross-validation folds. Each evaluated model therefore generated one independent prediction for each of the 500 EEG recordings. Average precision was used as the principal ranking measure because seizure recordings constituted 20% of the recording-level dataset. ROC-AUC, accuracy, balanced accuracy, sensitivity, specificity, F1-score, Matthews correlation coefficient, Brier score, and classification-error counts were used to provide complementary assessments of discrimination, classification, and probability quality.

4.1. Overall Recording-Level Performance

4.1.1. Comparative Performance of All Evaluated Models

Average precision (AP) was used as the primary ranking metric because seizure recordings represented only 20% of the recording-level dataset. Sensitivity and F1-score were treated as clinically relevant secondary measures because missed seizure events are particularly important in continuous and wearable monitoring applications [18]. Table 3 provides a common-reference comparison of the classical machine-learning, deep-learning, convolutional-kernel, and ensemble models.

The comparison included seven classical machine-learning models, seven neural and deep-learning models, three convolutional-kernel models, and one probability-level meta-ensemble. All 18 systems were ranked using recording-level AP obtained from the authoritative common-reference predictions. ROC-AUC and Brier score were used to evaluate discrimination and probability quality, whereas accuracy, balanced accuracy, sensitivity, specificity, F1-score, and MCC described performance at the selected classification threshold. Results based on family-specific aggregation rules are reported separately in Section 4.1.2 and Table 4.

Table 3. Recording-level performance of all evaluated models

Rank	Model	Family	AP	ROC-AUC	Accuracy	Balanced accuracy	Sensitivity	Specificity	F1-score	MCC	Brier score
1	RBF-SVM	Classical	0.997985	0.999475	0.990	0.98250	0.97	0.9950	0.974874	0.968651	0.008056
2	Extra Trees	Classical	0.996708	0.999150	0.988	0.98125	0.97	0.9925	0.970000	0.962500	0.009592
3	Gradient Boosting	Classical	0.996681	0.999150	0.986	0.97625	0.96	0.9925	0.964824	0.956104	0.008802
4	MultiRocket	Kernel	0.996591	0.998950	0.990	0.98250	0.97	0.9950	0.974874	0.968651	0.006004
5	Random Forest	Classical	0.996352	0.999025	0.986	0.97625	0.96	0.9925	0.964824	0.956104	0.010517
6	MLP	Deep	0.996129	0.998950	0.988	0.97750	0.96	0.9950	0.969697	0.962293	0.009542
7	k-NN	Classical	0.996107	0.999000	0.980	0.96500	0.94	0.9900	0.949495	0.937102	0.010155
8	InceptionTime	Deep	0.995953	0.999000	0.992	0.98375	0.97	0.9975	0.979798	0.974888	0.007003
9	ROCKET	Kernel	0.995153	0.998700	0.994	0.99250	0.99	0.9950	0.985075	0.981339	0.006147
10	TCN	Deep	0.994992	0.998800	0.994	0.99250	0.99	0.9950	0.985075	0.981339	0.007706
11	Logistic Regression	Classical	0.994559	0.998650	0.984	0.98250	0.98	0.9850	0.960784	0.951027	0.011111
12	CNN-1D	Deep	0.993024	0.997850	0.986	0.97250	0.95	0.9950	0.964467	0.955924	0.011176
13	MiniRocket	Kernel	0.992977	0.997925	0.992	0.99125	0.99	0.9925	0.980198	0.975261	0.007333
14	EEGNet-1D	Deep	0.990188	0.997675	0.984	0.98250	0.98	0.9850	0.960784	0.951027	0.014151
15	Gaussian Naive Bayes	Classical	0.985670	0.995825	0.974	0.95375	0.92	0.9875	0.934010	0.917991	0.019058
16	LSTM	Deep	0.979155	0.996050	0.986	0.98375	0.98	0.9875	0.965517	0.956901	0.015721
17	SimpleRNN	Deep	0.974147	0.993225	0.952	0.94000	0.92	0.9600	0.884615	0.855377	0.034077
18	Selected meta-ensemble	Ensemble	0.966604	0.988738	0.990	0.98250	0.97	0.9950	0.974874	0.968651	0.007065

Under the common-reference evaluation, RBF-SVM achieved the highest AP of 0.997985 and the highest ROC-AUC of 0.999475. It correctly classified 495 of the 500 recordings and obtained an MCC of 0.968651, making it the highest-ranked model according to the primary discrimination criterion.

Extra Trees and Gradient Boosting ranked second and third, respectively. MultiRocket ranked fourth and achieved the lowest Brier score among all evaluated models, indicating the smallest recording-level probability error under the common-reference configuration. The presence of several classical models near the top of the ranking demonstrates that the handcrafted EEG representation remained highly competitive under recording-grouped nested evaluation.

MLP was the highest-ranked neural model according to AP. InceptionTime and TCN achieved higher threshold-dependent F1-scores, but slightly lower AP values. This result demonstrates that strong performance at one selected operating threshold does not necessarily correspond to the best probability ranking across the complete precision–recall curve.

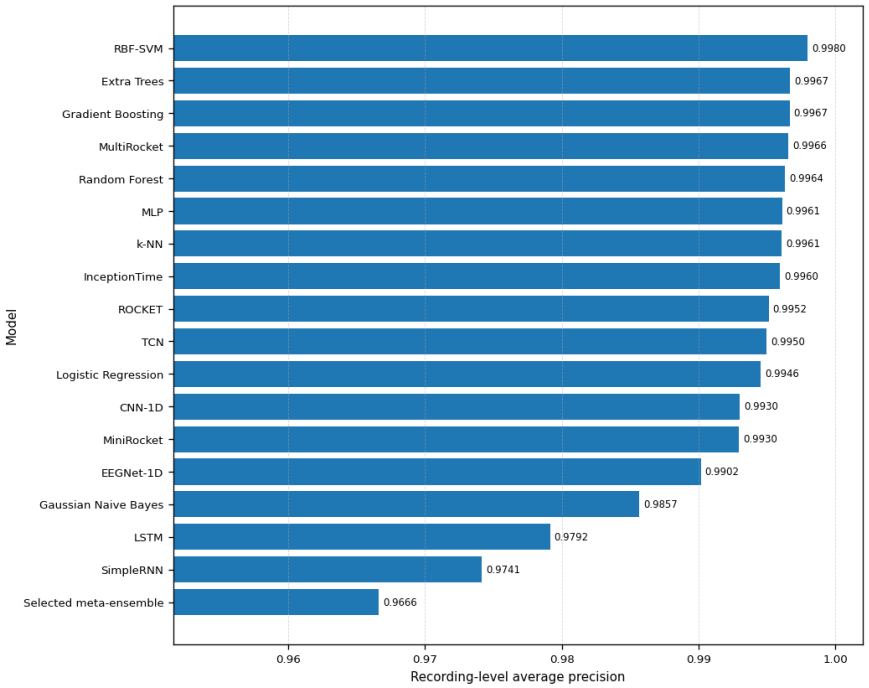


Fig. 1. Recording-level average precision across all evaluated models under grouped nested cross-validation.

The selected meta-ensemble achieved an accuracy of 0.990, specificity of 0.995, and MCC of 0.968651. However, its AP was lower than that of the individual family representatives. The ensemble therefore produced a strong threshold-specific operating point but did not improve the overall ranking of seizure and non-seizure recordings. The complete AP ranking is presented in Fig. 1.

4.1.2. Performance of the Model-Family Representatives

The strongest representatives of the classical, deep-learning, and convolutional-kernel families were RBF-SVM, MLP, and MultiRocket, respectively. The selected meta-ensemble represented the ensemble family. Recording-level aggregation was selected independently through grouped inner validation. Mean aggregation was selected for RBF-SVM, logit-mean aggregation for MLP, and median aggregation for MultiRocket. The ensemble used its fold-specific model-selection and probability-combination configuration. Table 4 presents the resulting recording-level performance.

Table 4. Recording-level performance of the model-family representatives using the selected aggregation rules

Family	Model	Aggregation	AP	ROC-AUC	Accuracy	Sensitivity	Specificity	F1-score	MCC	Brier score	Errors
Classical	RBF-SVM	Mean	0.997985	0.999475	0.990	0.97	0.9950	0.974874	0.968651	0.008056	5
Deep learning	MLP	Logit mean	0.997013	0.999200	0.986	0.96	0.9925	0.964824	0.956104	0.008070	7
Convolutional kernel	MultiRocket	Median	0.997366	0.999325	0.990	0.97	0.9950	0.974874	0.968651	0.006148	5
Ensemble	Selected meta-ensemble	Model-specific	0.966604	0.988738	0.990	0.97	0.9950	0.974874	0.968651	0.007065	5

RBF-SVM remained the highest-ranked family representative, achieving an AP of 0.997985 and a ROC-AUC of 0.999475. MultiRocket produced a slightly lower AP of 0.997366 but achieved the lowest Brier score of 0.006148, indicating the most accurate probability estimates among the family representatives. Logit-mean aggregation improved the ranking performance of MLP relative to arithmetic-mean aggregation, although its threshold-dependent MCC and accuracy remained slightly lower than those of RBF-SVM and MultiRocket.

The precision–recall curves are presented in Fig. 2. RBF-SVM, MLP, and MultiRocket remained close to the upper region of the curve, whereas the meta-ensemble showed weaker precision–recall ordering.

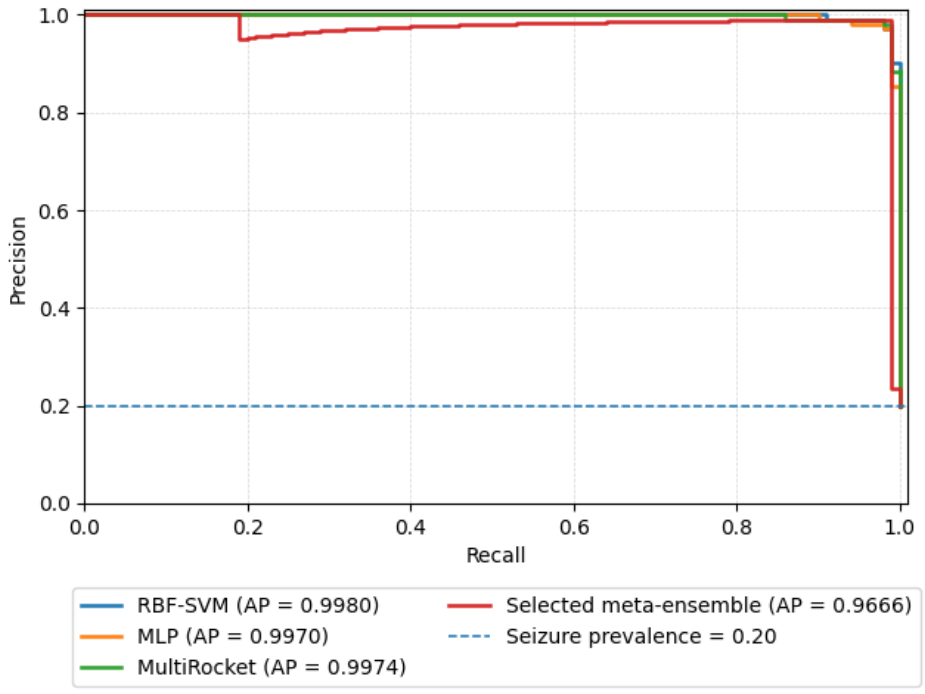


Fig. 2. Recording-level precision–recall curves of the model-family representatives using their selected aggregation rules.

The corresponding ROC curves are presented in Fig. 3. All four representatives achieved high ROC-AUC values, with RBF-SVM producing the highest point estimate.

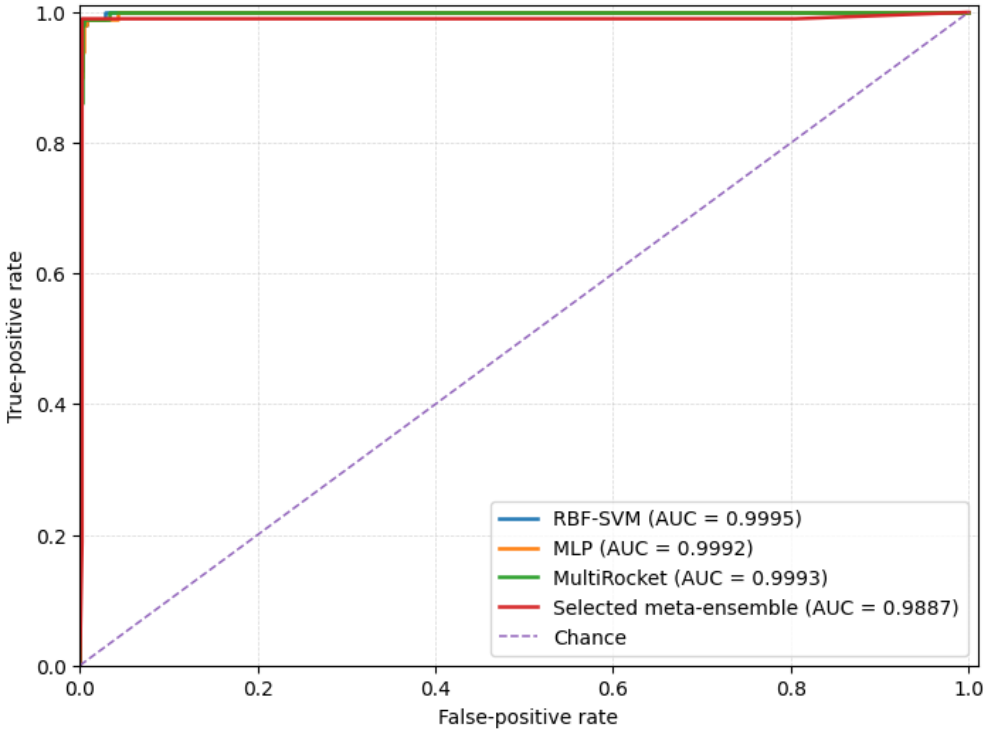


Fig. 3. Recording-level receiver operating characteristic curves of the model-family representatives using their selected aggregation rules.

4.2. Statistical Reliability and Probability Calibration

4.2.1. Bootstrap Confidence Intervals and Statistical Comparisons

The uncertainty of the family-representative performance estimates was evaluated using 2,000 stratified bootstrap repetitions. Resampling was conducted at the recording level while preserving the seizure and non-seizure class proportions. Table 5 presents the observed values and 95% bootstrap confidence intervals for average precision, ROC-AUC, MCC, and Brier score.

Table 5. Bootstrap confidence intervals for the model-family representatives

Model	AP [95% CI]	ROC-AUC [95% CI]	MCC [95% CI]	Brier score [95% CI]
RBF-SVM	0.997985 [0.994372, 1.000000]	0.999475 [0.998500, 1.000000]	0.968651 [0.937500, 0.993746]	0.008056 [0.003779, 0.013455]
MLP	0.997013 [0.992305, 1.000000]	0.999200 [0.997825, 1.000000]	0.956104 [0.923909, 0.987484]	0.008070 [0.002617, 0.014957]
MultiRocket	0.997366 [0.992676, 1.000000]	0.999325 [0.998100, 1.000000]	0.968651 [0.937102, 0.993746]	0.006148 [0.001059, 0.013250]
Selected meta-ensemble	0.966604 [0.908713, 1.000000]	0.988738 [0.968287, 1.000000]	0.968651 [0.937500, 0.993746]	0.007065 [0.002310, 0.013682]

RBF-SVM produced the highest observed AP and ROC-AUC, whereas MultiRocket produced the lowest observed Brier score. However, paired bootstrap comparisons identified no statistically significant differences among RBF-SVM, MLP, MultiRocket, and the selected meta-ensemble for AP, ROC-AUC, MCC, or Brier score after Holm correction. Accordingly, the numerical ranking should be interpreted as an ordering of point estimates rather than evidence of statistically established superiority.

Exact McNemar tests likewise found no significant differences in recording-level error patterns among the four family representatives after Holm correction. Although MLP produced seven errors, compared with five errors for each of the other representatives, these differences were not statistically resolved.

The paired-bootstrap effect estimates and corresponding 95% confidence intervals are presented in Fig. 4. Confidence intervals crossing zero indicate that the corresponding pairwise differences were not statistically resolved.

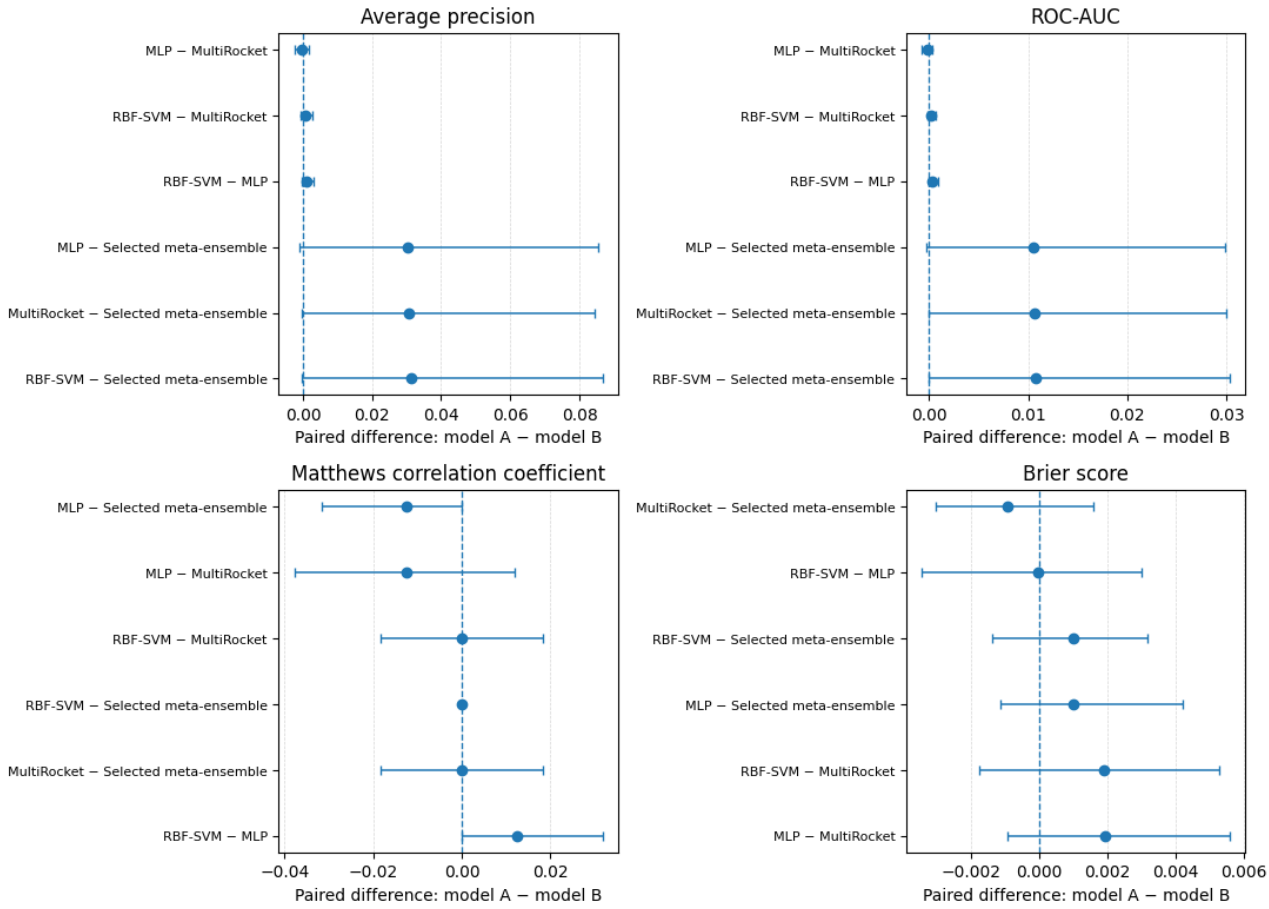


Fig. 4. Paired bootstrap differences and 95% confidence intervals among the model-family representatives.

4.2.2. Probability Calibration

Average precision and ROC-AUC evaluate discrimination but do not determine whether predicted probabilities correspond closely to observed recording outcomes. Probability accuracy was therefore assessed using the Brier score and recording-level calibration curves. Lower Brier scores indicate smaller squared differences between predicted probabilities and actual class labels.

MultiRocket produced the lowest family-representative Brier-score point estimate of 0.006148 under median aggregation. The selected meta-ensemble followed with 0.007065, while RBF-SVM and MLP produced values of 0.008056 and 0.008070, respectively. The bootstrap confidence intervals overlapped, and the paired differences among the four representatives were not statistically significant after Holm correction. MultiRocket should therefore be described as having the lowest observed probability error, rather than statistically superior calibration.

These findings demonstrate that the model with the highest discrimination was not necessarily the model with the lowest probability-error estimate. The recording-level calibration curves, generated using the selected aggregation rule for each representative, are presented in Fig. 5.

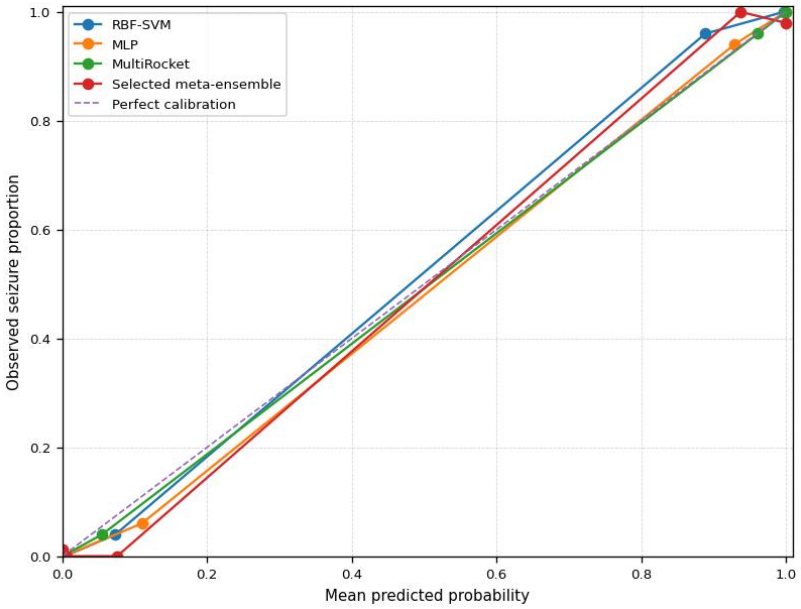


Fig. 5. Recording-level calibration curves of the model-family representatives using their selected aggregation rules.

4.3. Misclassification Behaviour

4.3.1. Recording-Level Confusion Matrices

The classification-error analysis was performed using the pooled outer-fold decisions of all 18 systems. Each model was evaluated on the same 400 non-seizure and 100 seizure recordings. This allowed false-positive and false-negative behaviour to be compared directly without differences in sample composition. The complete confusion-matrix values are reported in Table 6.

Table 6. Recording-level confusion matrices of all evaluated models

Model	TN	FP	FN	TP	Sensitivity	Specificity	Precision	F1-score	MCC
RBF-SVM	398	2	3	97	0.97	0.9950	0.979798	0.974874	0.968651
Extra Trees	397	3	3	97	0.97	0.9925	0.970000	0.970000	0.962500
Gradient Boosting	397	3	4	96	0.96	0.9925	0.969697	0.964824	0.956104
MultiRocket	398	2	3	97	0.97	0.9950	0.979798	0.974874	0.968651
Random Forest	397	3	4	96	0.96	0.9925	0.969697	0.964824	0.956104
MLP	398	2	4	96	0.96	0.9950	0.979592	0.969697	0.962293
k-NN	396	4	6	94	0.94	0.9900	0.959184	0.949495	0.937102
InceptionTime	399	1	3	97	0.97	0.9975	0.989796	0.979798	0.974888
ROCKET	398	2	1	99	0.99	0.9950	0.980198	0.985075	0.981339
TCN	398	2	1	99	0.99	0.9950	0.980198	0.985075	0.981339
Logistic Regression	394	6	2	98	0.98	0.9850	0.942308	0.960784	0.951027
CNN-1D	398	2	5	95	0.95	0.9950	0.979381	0.964467	0.955924
MiniRocket	397	3	1	99	0.99	0.9925	0.970588	0.980198	0.975261
EEGNet-1D	394	6	2	98	0.98	0.9850	0.942308	0.960784	0.951027
Gaussian Naive Bayes	395	5	8	92	0.92	0.9875	0.948454	0.934010	0.917991
LSTM	395	5	2	98	0.98	0.9875	0.951456	0.965517	0.956901
SimpleRNN	384	16	8	92	0.92	0.9600	0.851852	0.884615	0.855377
Selected meta-ensemble	398	2	3	97	0.97	0.9950	0.979798	0.974874	0.968651

ROCKET and TCN produced the lowest total error count, with two false positives and one false negative each. MiniRocket also missed only one seizure recording but generated three false-positive decisions. RBF-SVM, MultiRocket, and the ensemble each produced two false positives and three false negatives.

InceptionTime produced only one false-positive recording and achieved the highest specificity among the deep-learning models. SimpleRNN produced the largest number of errors, with 16 false positives and 8 false negatives. Gaussian Naive Bayes also showed reduced seizure sensitivity, missing eight seizure recordings. The confusion matrices of the family representatives are presented in Fig. 6.

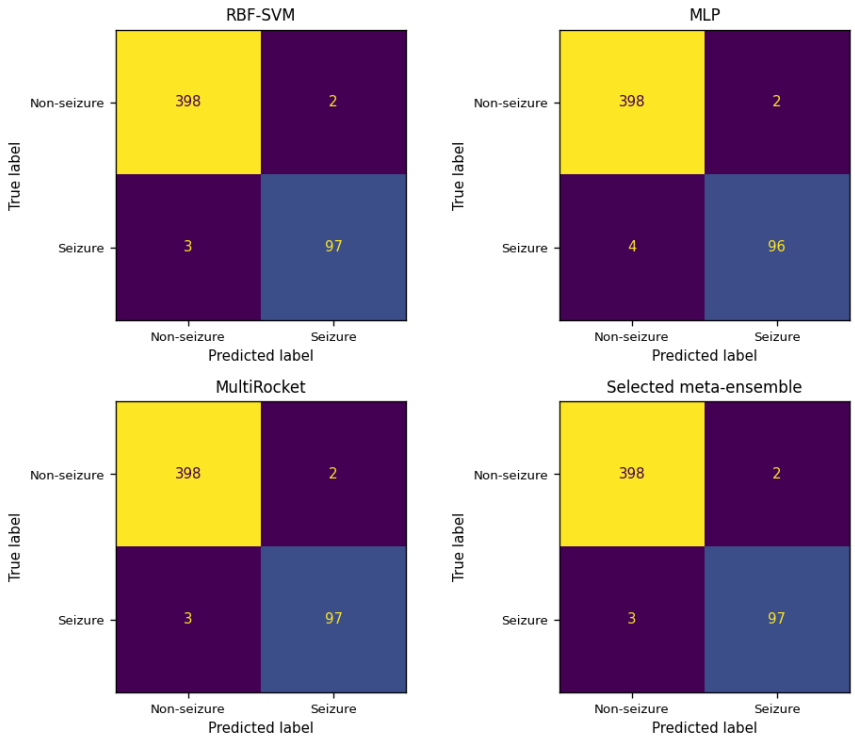


Fig. 6. Recording-level confusion matrices of the family-champion models.

4.3.2. Distribution of False-Positive and False-Negative Errors

The error-based ranking was not identical to the average-precision ranking. ROCKET and TCN produced fewer thresholded classification errors than RBF-SVM, although RBF-SVM achieved higher average precision. This difference occurred because the confusion matrix evaluates one selected operating threshold, whereas average precision evaluates probability ordering across the complete range of recall values.

False-negative counts ranged from one for ROCKET, TCN, and MiniRocket to eight for Gaussian Naive Bayes and SimpleRNN. False-positive counts ranged from one for InceptionTime to 16 for SimpleRNN. The lower MCC of SimpleRNN resulted from its comparatively high number of errors in both classes. The distribution of false-positive and false-negative errors is presented in Fig. 7.

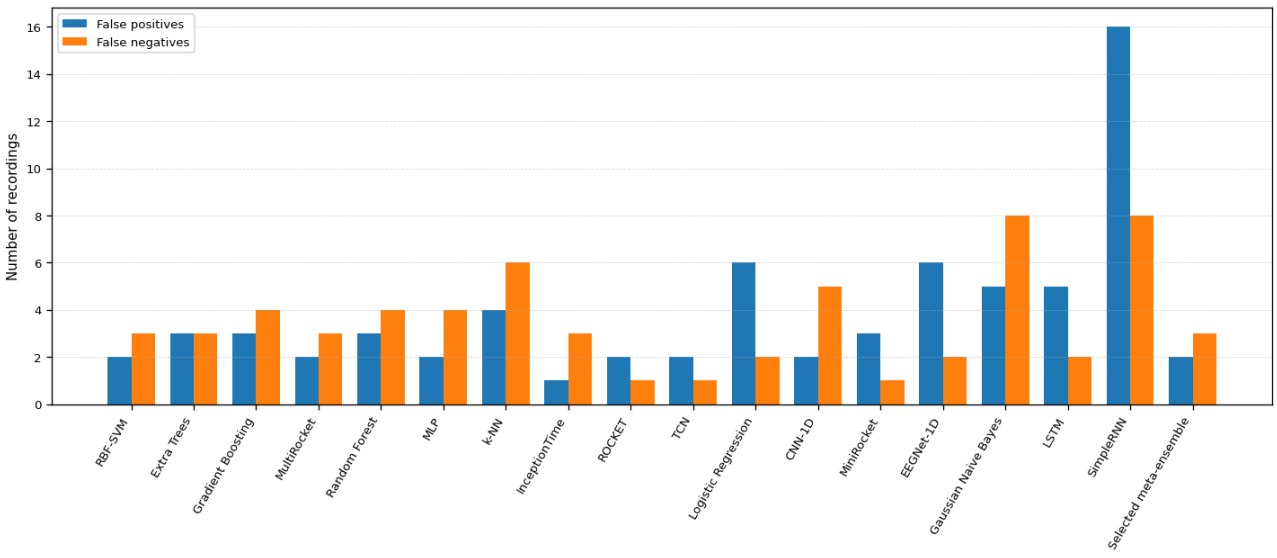


Fig. 7. False-positive and false-negative counts across all evaluated models.

4.4. Ablation Analysis

4.4.1. Feature-Domain and Preprocessing Ablation

A structured ablation analysis was conducted to examine the contributions of the three handcrafted feature domains and the selected scaling procedure to RBF-SVM performance. Nine configurations were evaluated across the five outer folds, including individual feature domains, leave-one-domain-out representations, robust scaling, and no scaling. The same recording assignments, grouped evaluation protocol, and RBF-SVM configuration were retained across all experiments. All preprocessing parameters were estimated exclusively from the corresponding training partitions, and performance was calculated from the pooled outer-test recording predictions. Table 7 presents the resulting feature-domain and preprocessing-ablation performance.

Table 7. Feature-domain and preprocessing ablation results

Rank	Feature configuration	Number of features	Scaling	AP	ROC-AUC	MCC	Brier score	Errors
1	All feature domains	67	Standard	0.997985	0.999475	0.968651	**0.008056**	5
2	Without frequency-domain features	37	Standard	0.997282	0.999300	0.962500	0.008978	6
3	Without time-domain features	39	Standard	0.997241	0.999275	0.956104	0.009333	7
4	Without nonlinear features	58	Standard	0.997235	0.999200	0.968569	0.008601	5
5	Time-domain features only	28	Standard	0.997100	0.999250	0.962500	0.010156	6
6	Frequency-domain features only	30	Standard	0.996340	0.999025	0.955924	0.010550	7
7	All feature domains	67	Robust	0.994870	0.998550	0.949697	0.011181	8
8	Nonlinear features only	9	Standard	0.992307	0.996025	0.930635	0.014426	11
9	All feature domains	67	None	0.850775	0.963700	0.608959	0.054383	92

The complete standardised 67-feature representation achieved the highest AP, ROC-AUC, and MCC. Removing any one feature domain produced an AP reduction of less than 0.001, indicating that predictive information was distributed across the time-domain, frequency-domain, and nonlinear feature groups rather than being attributable to a single domain.

The time-domain-only representation remained highly competitive, achieving an AP of 0.997100 and six recording-level errors. The frequency-domain-only representation also retained strong discrimination, whereas the nonlinear-only representation produced the weakest result among the standardised feature subsets. Nevertheless, the nonlinear-only configuration still achieved an AP above 0.99.

The scaling procedure had a substantially greater effect than the removal of an individual feature domain. Robust scaling produced a moderate reduction in performance compared with standard scaling. In contrast, omitting scaling reduced AP from 0.997985 to 0.850775 and MCC from 0.968651 to 0.608959, while increasing the number of errors from 5 to 92. In one outer fold, the unscaled configuration produced an AP of 0.20 and an MCC of 0.00. These results demonstrate that fold-specific standardisation was a critical component of the RBF-SVM pipeline. The feature-domain and preprocessing-ablation results are presented in Fig. 8.

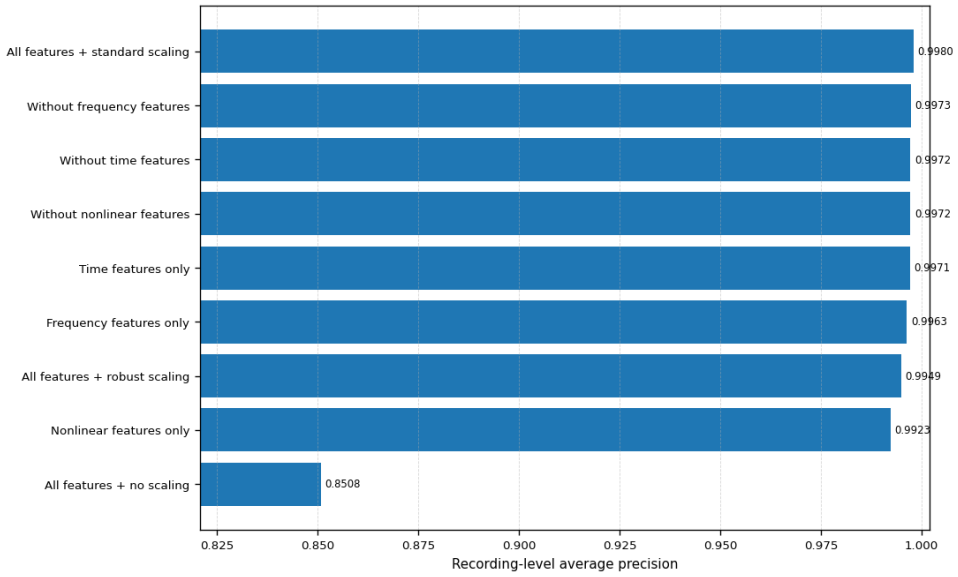


Fig. 8. Feature-domain and preprocessing ablation results for the reference RBF-SVM.

4.4.2. Recording-Level Probability-Aggregation Ablation

The aggregation analysis examined how the 23 segment probabilities belonging to each recording should be combined into one recording-level probability. Eight aggregation functions were evaluated for RBF-SVM, MLP, and MultiRocket. Selection was based exclusively on inner recording-level predictions, while the outer folds were used only for final evaluation. Table 8 compares the arithmetic-mean reference with the selected aggregation for each model family.

Table 8. Effect of the selected recording-level aggregation rules

Model	Reference aggregation	Reference AP	Selected aggregation	Selected AP	AP change	Selected ROC-AUC	Selected MCC	Selected Brier score
RBF-SVM	Mean	0.997985	Mean	0.997985	0.000000	0.999475	0.968651	0.008056
MLP	Mean	0.996129	Logit mean	0.997013	+0.000884	0.999200	0.956104	0.008070
MultiRocket	Mean	0.996591	Median	0.997366	+0.000775	0.999325	0.968651	0.006148

Mean aggregation remained optimal for RBF-SVM. Logit-mean aggregation improved MLP average precision from 0.996129 to 0.997013 and reduced its Brier score from 0.009542 to 0.008070. Median aggregation improved MultiRocket average precision from 0.996591 to 0.997366.

Maximum aggregation was less stable. For RBF-SVM, it reduced average precision to 0.988361 and increased the total error count to 11. Maximum aggregation also produced substantially larger Brier scores for MLP and MultiRocket. This finding indicates that a decision based primarily on the most seizure-like segment was less reliable than an aggregation that considered the complete recording. The aggregation-ablation results are presented in Fig. 9.

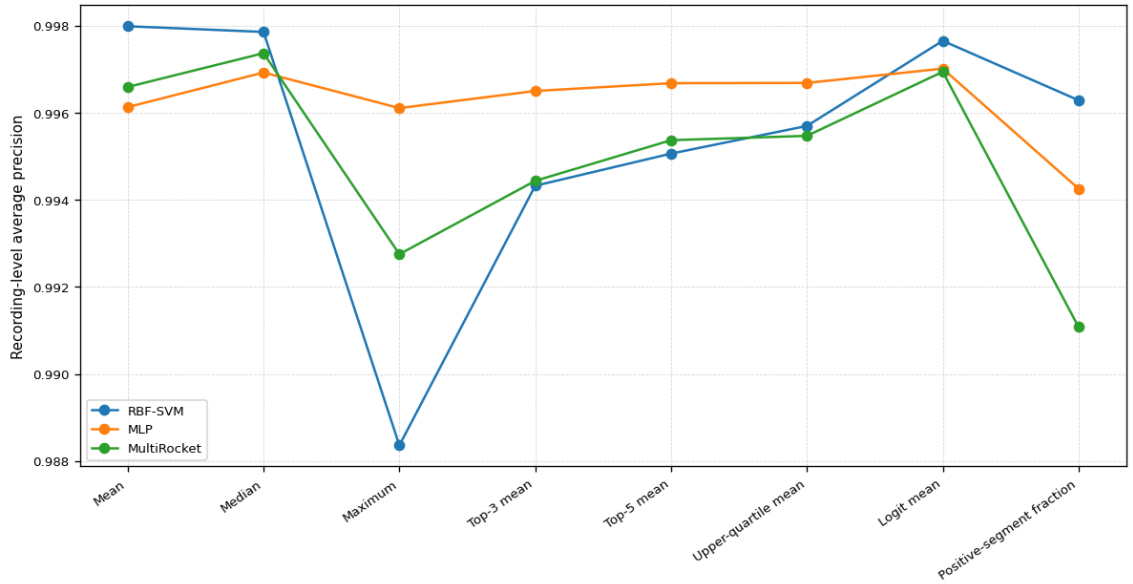


Fig. 9. Recording-level probability-aggregation ablation for the classical, deep-learning, and convolutional-kernel family champions.

4.4.3. Kernel-Transformation and Classifier Ablation

The convolutional-kernel ablation examined the effects of transformation type, kernel budget, transformed-feature dimensionality, downstream feature selection, Logistic Regression regularisation, and class weighting. Small, medium, and large kernel-budget configurations were compared using grouped inner-fold recording-level average precision. The pooled outer-fold results of the predefined classifier variants were then used only to report their final performance. Table 9 summarises the principal findings.

Table 9. Kernel-transformation and classifier-ablation results

Transformation	Best feature-budget tier	Mean inner AP	Highest-performing classifier variant	Pooled AP	MCC	Brier score
MiniRocket	Small	0.995309	Balanced, (C=1), with 2,000 selected features	0.996006	0.981339	0.007525
MultiRocket	Small	0.997261	Balanced, (C=10), using all features	0.996947	0.981199	0.005894
ROCKET	Small	0.995327	Balanced, (C=1), with 2,000 selected features	0.996764	0.975261	0.006536

Note: Mean inner AP was used to compare the kernel-budget tiers. Pooled AP, MCC, and Brier score represent outer-fold results for the corresponding predefined classifier-ablation variants.

Increasing the kernel budget did not consistently improve inner recording-level AP. For MultiRocket, the small transformation produced a mean inner AP of 0.997261, compared with 0.996231 and 0.996458 for the medium and large configurations, respectively. Comparable patterns were observed for MiniRocket and ROCKET, indicating that a larger transformation did not necessarily provide better generalisation.

The downstream Logistic Regression configuration also influenced performance. For MiniRocket and ROCKET, balanced classification with $C = 1$ and selection of 2,000 transformed features produced pooled AP values of 0.996006 and 0.996764, respectively. For MultiRocket, balanced Logistic Regression with $C = 10$using all transformed features produced the lowest Brier score in the ablation analysis, at 0.005894.

A separate MultiRocket variant using 2,000 selected features achieved an accuracy of 0.996 and an MCC of 0.987500. This variant was evaluated as part of the classifier-ablation analysis and was not the final MultiRocket family representative reported in Table 4.

The MultiRocket results in Tables 4 and 9 correspond to different experimental stages. Table 9 reports predefined kernel and classifier-ablation variants, whereas Table 4 reports the final family representative using median recording-level aggregation selected through grouped inner validation. Consequently, the Table 9 pooled AP of 0.996947 should not be expected to equal the Table 4 AP of 0.997366.

Overall, the results show that increasing the number of convolutional features did not automatically improve recording-level performance. The interaction among transformation type, feature dimensionality, class weighting, Logistic Regression regularisation, and recording-level aggregation was more influential than kernel budget alone. The complete ablation results are presented in Fig. 10.

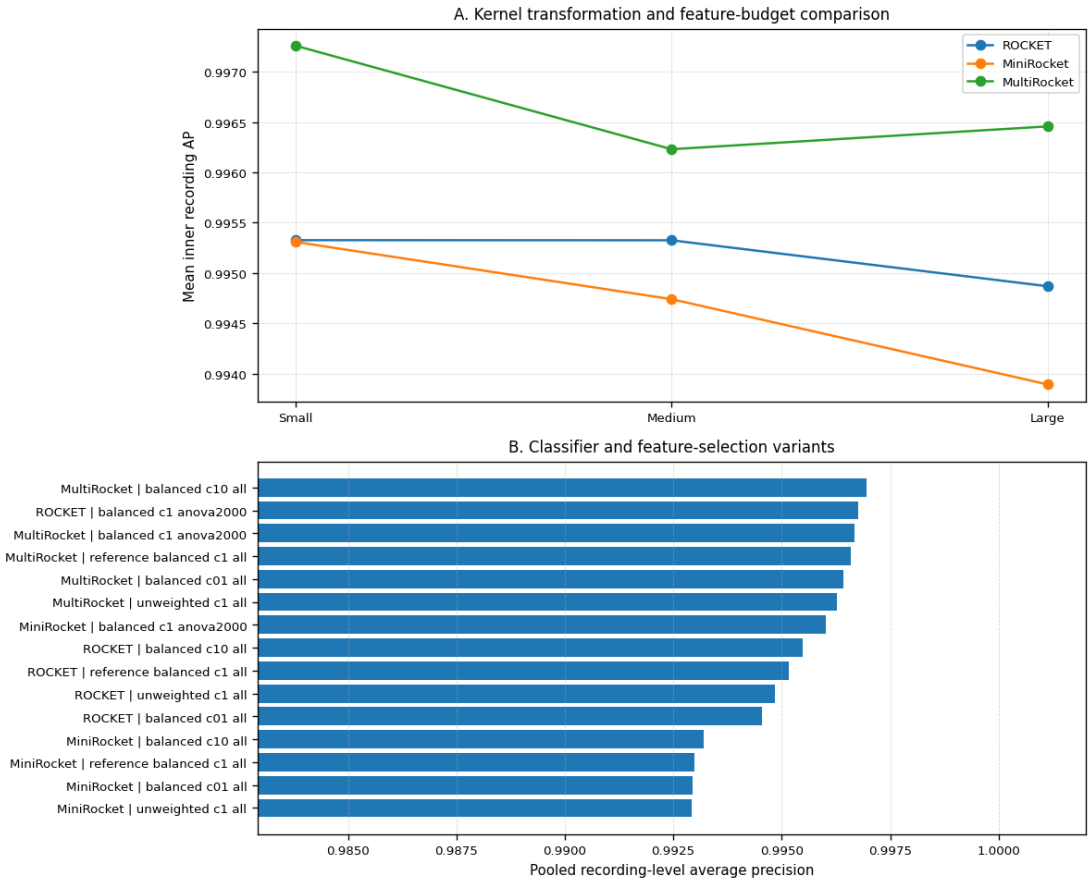


Fig. 10. Kernel-transformation, feature-budget, and classifier-ablation results for ROCKET, MiniRocket, and MultiRocket.

4.5. Explainability Results

4.5.1. RBF-SVM Feature Importance

Recording-block permutation importance was used to determine which handcrafted EEG features contributed most strongly to RBF-SVM predictions. All 23 segments belonging to one recording were treated as a block during permutation. The analysis covered 67 features, five outer folds, and ten repetitions per feature, resulting in 3,350 permutation evaluations. Table 10 presents the ten highest-ranked features according to the mean decrease in recording-level average precision.

Table 10. Leading recording-block permutation features for RBF-SVM

Rank	Feature	Domain	Mean AP decrease
1	Approximate entropy	Nonlinear	0.001672
2	Low-to-high gamma power ratio	Frequency	0.001532
3	Signal mean	Time	0.001361
4	Theta relative power	Frequency	0.000822
5	Mean absolute value	Time	0.000771
6	Hurst exponent	Nonlinear	0.000724
7	Mean absolute difference	Time	0.000697
8	Line length	Time	0.000687
9	Crest factor	Time	0.000671
10	Median absolute deviation	Time	0.000650

Approximate entropy was the highest-ranked feature, followed by the low-to-high gamma power ratio and signal mean. The leading variables represented all three feature domains, which was consistent with the ablation finding that no single domain completely explained the classifier performance.

The individual AP decreases were relatively small because RBF-SVM operated close to the upper performance boundary and several EEG features contained correlated information. The importance values therefore represent model-specific predictive contributions rather than independent causal effects. The leading feature-level results are presented in Fig. 11.

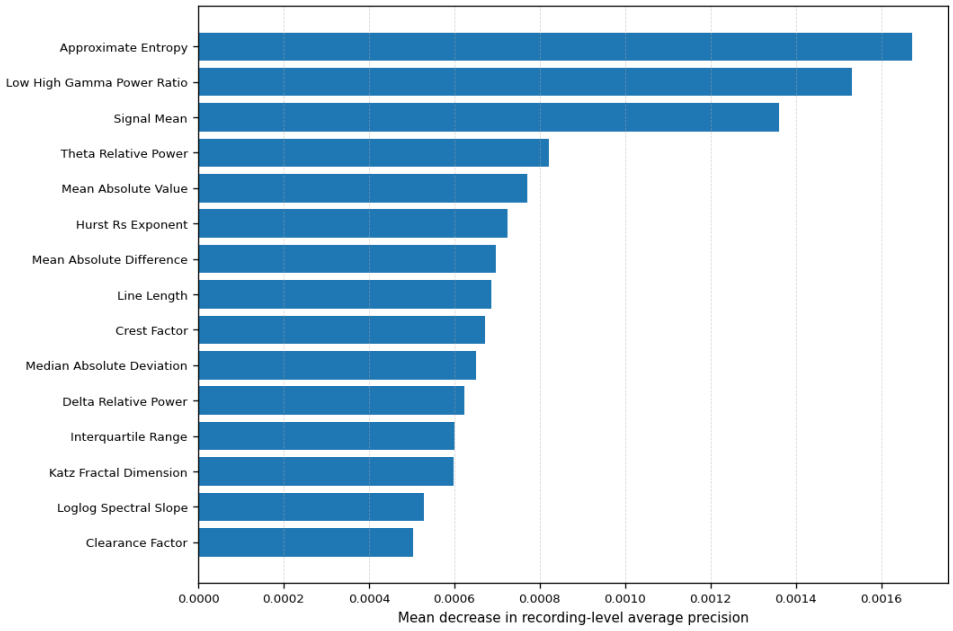


Fig. 11. Recording-block permutation importance of the leading handcrafted EEG features for RBF-SVM.

At the domain level, nonlinear and time-domain features produced similar mean decreases in average precision, whereas frequency-domain features showed a lower mean effect. Nevertheless, the frequency-domain group contained the second- and fourth-ranked individual features. The domain-level results are shown in Fig. 12.

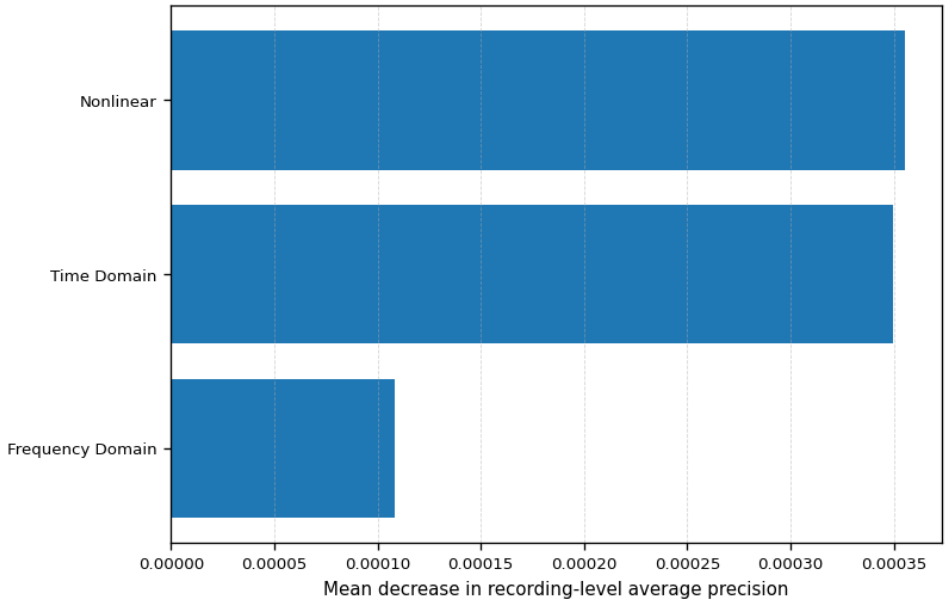


Fig. 12. Feature-domain permutation importance for the reference RBF-SVM.

4.5.2. Neural-Model Attribution Analysis

The attribution analysis covered seven neural architectures, five outer folds, and 20 selected recordings per model and fold, resulting in 700 model–recording evaluations. The selected cases included both correctly and incorrectly classified recordings, with 59 misclassified cases represented in the analysis.

Attribution faithfulness was evaluated by masking the 10%, 20%, and 30% most highly attributed inputs and comparing the resulting reduction in prediction confidence with that produced by random masking at the same proportions. For the MLP, masking was applied to handcrafted feature dimensions, whereas for CNN-1D, SimpleRNN, LSTM, TCN, InceptionTime, and EEGNet-1D, masking was applied to temporal positions in the raw EEG sequence. A positive faithfulness advantage indicates that masking highly attributed inputs reduced prediction confidence more strongly than random masking. Table 11 reports the mean faithfulness advantage at the 20% masking level and the mean fold-wise Spearman correlation of the global attribution rankings.

Table 11. Neural-model attribution faithfulness and fold stability

Model	Faithfulness advantage at 20% masking	Mean Spearman correlation
MLP	0.068066	0.570820
CNN-1D	−0.035694	0.182482
SimpleRNN	0.138428	0.230170
LSTM	0.155971	0.685562
TCN	0.005878	0.161424
InceptionTime	0.032502	0.193764
EEGNet-1D	0.036527	0.610162

LSTM achieved the largest mean faithfulness advantage at the 20% masking level and the highest mean fold-wise attribution stability. This indicates that its most highly attributed temporal positions generally contributed meaningfully to prediction confidence and that its global attribution rankings were comparatively consistent across outer folds.

SimpleRNN also produced a relatively large faithfulness advantage. However, its lower mean Spearman correlation indicates that the temporal positions identified as important varied considerably across folds. EEGNet-1D produced the second-highest attribution stability, although its masking advantage was smaller than those of LSTM and SimpleRNN.

The MLP demonstrated a positive faithfulness advantage and moderate fold stability. Its leading handcrafted inputs included the low-to-high gamma power ratio, approximate entropy, theta relative power, Hjorth complexity, mean absolute difference, and line length. These findings were consistent with the RBF-SVM permutation analysis, which also identified contributions from time-domain, frequency-domain, and nonlinear features.

CNN-1D produced a negative mean faithfulness advantage at the 20% masking level. Thus, masking its most highly attributed temporal positions did not, on average, reduce confidence more strongly than random masking. TCN and InceptionTime also showed relatively small faithfulness advantages and low fold-wise stability. These results indicate

that attribution reliability differed substantially across neural architectures and that high predictive performance did not necessarily correspond to stable or faithful explanations.

The masking-based faithfulness results across the evaluated masking proportions are presented in Fig. 13, while the fold-wise stability of the global attribution rankings is shown in Fig. 14.

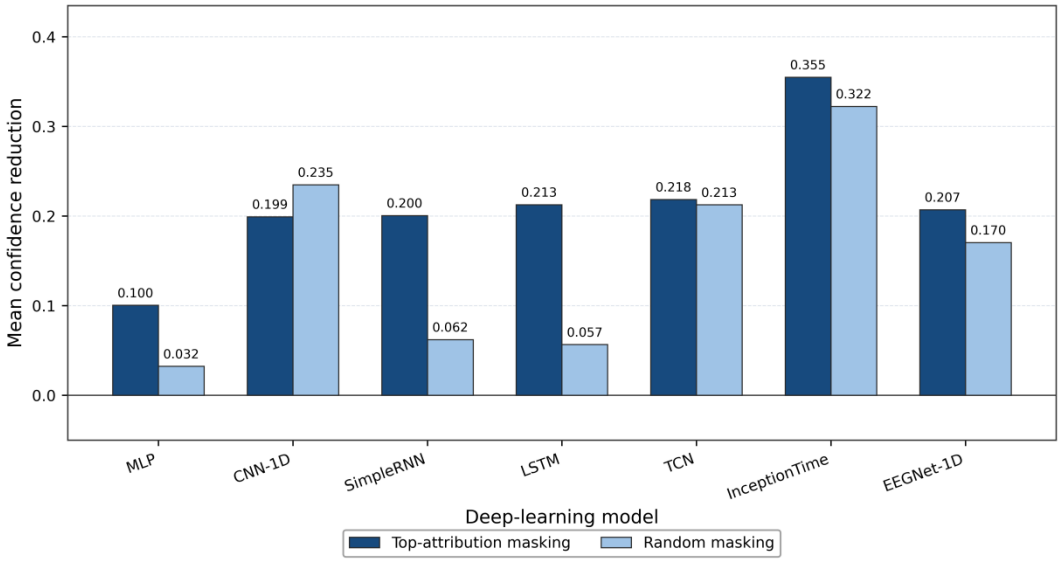


Fig. 13. Neural-model attribution faithfulness under top-attribution and random input masking.

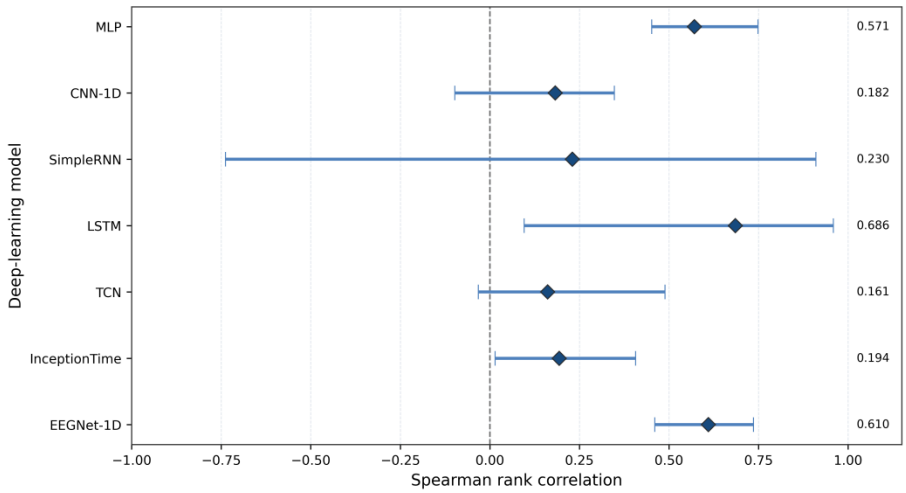


Fig. 14. Fold-wise stability of neural-model attribution rankings measured using Spearman correlation.

4.6. Computational Efficiency

4.6.1. Inference Latency, Memory, and Artifact Size

Computational efficiency was evaluated using the saved outer-fold artifacts of RBF-SVM, MLP, MultiRocket, and the selected meta-ensemble. Each timing run processed the complete outer-test partition, comprising 100 recordings and 2,300 segment-level inputs. Three warm-up runs were performed before 20 timed repetitions for each model and outer fold. Artifact-loading time was measured separately and was excluded from inference latency.

For RBF-SVM and MLP, latency was measured from the available 67-dimensional handcrafted-feature matrix and included fold-specific standardisation, segment-level probability generation, and recording-level aggregation. The time required to extract the 67 features from raw EEG was not included. For MultiRocket, latency was measured from the raw 178-sample EEG sequences and included signal normalisation, the MultiRocket transformation, transformed-feature standardisation, Logistic Regression inference, and median recording-level aggregation. The ensemble benchmark started from the 100×17matrix of previously generated base-model probabilities and measured only probability weighting and optional calibration. Table 12 presents the resulting efficiency measurements.

Table 12. Computational efficiency of the model-family representatives

Model	Aggregation	Median full-fold latency (ms)	Mean full-fold latency (ms)	Median throughput (inputs/s)	Mean Python peak memory (bytes)	Median loading time (ms)	Median artifact size (bytes)
Ensemble combination	Model-specific probabilities	0.0069	0.01310	14,492,750	3,487	1.3389	1,325
MLP	Logit mean	4.3971	4.66520	523,072	1,880,618	5.5906	75,269
RBF-SVM	Mean	63.4022	61.48826	36,276	2,535,414	5.5290	275,524
MultiRocket	Median	881.5317	611.55220	2,609	413,230,406	0.7659	667,926

Note: For RBF-SVM, MLP, and MultiRocket, throughput represents EEG segments processed per second. For the ensemble, it represents recording-level probability vectors combined per second. Ensemble throughput is therefore not directly comparable with standalone-model throughput.

MLP provided the lowest standalone inference latency, requiring a median of 4.3971 ms to process one complete outer-test fold. It also had the smallest standalone serialized artifact. However, this value represents inference after handcrafted-feature extraction and should not be interpreted as complete processing from raw EEG.

RBF-SVM required a median of 63.4022 ms per complete outer-test fold. Although it was slower than MLP, it achieved the highest observed recording-level average precision. As with MLP, its latency included standardisation, prediction, and recording-level aggregation but excluded the initial extraction of handcrafted EEG features.

MultiRocket had the highest latency and memory requirements. Its median full-fold latency was 881.5317 ms, and its mean Python peak memory was approximately 413.23 million bytes. Its mean resident-memory increase was approximately 1.16 billion bytes. These requirements arose primarily from the convolutional transformation and the generation of up to 20,160 transformed features. Unlike the feature-based models, the MultiRocket benchmark included transformation of the raw EEG sequence, so its latency represents a broader portion of the inference pipeline.

The ensemble combination required only 0.0069 ms because it operated on base-model probabilities that had already been generated. This value excluded signal preprocessing, handcrafted-feature extraction, convolutional transformation, artifact loading, and inference by the 17 base models. It therefore represents incremental meta-ensemble overhead rather than complete end-to-end ensemble latency and should not be compared directly with the standalone-model timings.

4.6.2. Performance–Efficiency Trade-Off

The efficiency results show that no single system was optimal for every criterion. MLP provided the lowest standalone latency, RBF-SVM achieved the highest observed average precision, and MultiRocket produced the lowest family-representative Brier score. The preferred model therefore depends on whether discrimination, probability calibration, latency, or memory consumption is given the greatest importance.

The performance–efficiency relationship is presented in Fig. 15. MLP occupied the low-latency region while maintaining high average precision. RBF-SVM achieved the highest average precision at moderate latency, whereas MultiRocket required substantially greater computational resources for a relatively small difference in discrimination.

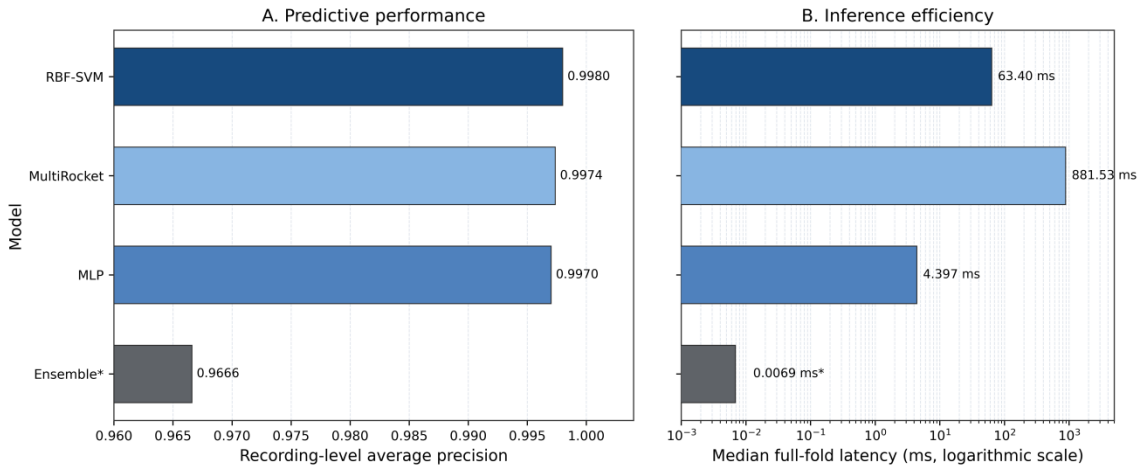


Fig. 15. Recording-level performance–efficiency trade-off of the family-champion models.

4.7. Integrated Discussion

4.7.1. Interpretation of the Comparative Findings

The comparative findings demonstrate that EEG seizure-recognition performance depended on the combined effects of signal representation, preprocessing, model structure, recording-level aggregation, and threshold selection. The

strongest observed system was RBF-SVM trained on the 67 handcrafted EEG features rather than a raw-sequence neural architecture.

The ablation findings showed that the RBF-SVM result was not attributable to one isolated feature domain. Removal of the time-domain, frequency-domain, or nonlinear group produced only modest reductions in performance. In contrast, omission of scaling caused a substantial decline, showing that fold-specific standardisation was a critical component of the final pipeline.

MLP was the highest-ranked neural model according to the common-reference average-precision comparison and remained highly competitive after logit-mean aggregation. Because MLP operated on the handcrafted representation, its result indicates that the quality of the input representation was at least as important as explicit convolutional or recurrent temporal modelling for this benchmark.

MultiRocket produced the lowest probability error among the family representatives and remained close to RBF-SVM in average precision. However, its computational requirements were considerably higher. It therefore offered a strong discrimination–calibration profile but a less favourable efficiency profile.

The ensemble did not improve average precision over the strongest individual models. Its high accuracy, MCC, and specificity indicate that the selected threshold generated effective binary decisions. However, the lower average precision suggests that combining correlated base-model probabilities did not improve the global ranking of seizure and non-seizure recordings.

4.7.2. Limitations and Future Work

The evaluation was recording-disjoint but cannot be described as patient-independent because verified patient identifiers were unavailable. The grouped protocol prevented segments from the same reconstructed recording from appearing in different folds, but it could not determine whether separate recordings originated from the same individual.

The dataset consisted of preprocessed, fixed-length, single-channel EEG segments. It did not reproduce the full complexity of raw clinical EEG, including multichannel montages, electrode artefacts, motion interference, long-duration non-stationarity, and variability among acquisition systems.

All models were developed and evaluated using one benchmark dataset. External generalisation to independent datasets, institutions, or acquisition systems was therefore not established. The binary formulation also combined four distinct non-seizure conditions into one category and did not evaluate multiclass physiological differentiation.

Although the dataset contained 11,500 segments, the independent recording-level evaluation consisted of 500 source recordings. The uncertainty estimates and statistical conclusions should consequently be interpreted at the recording level rather than the segment level.

The explainability analyses identified the features and input positions used by the fitted models, but they did not establish causal neurological mechanisms or clinically validated seizure biomarkers. Attribution stability also varied among neural architectures, indicating that explanations may depend on the model and data partition.

Future investigations should evaluate the framework using datasets with verified patient identities, raw multichannel EEG, heterogeneous acquisition conditions, and independent external test cohorts. Additional work should consider multiclass EEG-state classification, artefact-robust preprocessing, prospective calibration, clinically reviewed explanations, and complete end-to-end latency measurements that include signal preprocessing and feature generation.

5. Conclusion

This study presented a unified, leakage-controlled comparison of classical machine-learning, neural and deep-learning, convolutional-kernel time-series, and ensemble models for EEG-based seizure detection. Recording-grouped nested cross-validation was used to prevent segments from the same recording from appearing across training and test partitions. RBF-SVM achieved the highest recording-level average precision of 0.997985 and ROC-AUC of 0.999475, demonstrating the continued value of properly engineered time-, frequency-, and nonlinear-domain EEG features. MultiRocket achieved a slightly lower average precision but produced the lowest family-representative Brier score of 0.006148, indicating the smallest observed probability error. MLP provided the lowest standalone inference latency after feature extraction, while the selected meta-ensemble achieved strong threshold-dependent performance without improving overall precision–recall ranking. Paired bootstrap analysis found no statistically significant differences among the family representatives after multiple-comparison correction; therefore, their ranking should be interpreted as an ordering of observed point estimates rather than proof of universal superiority. Ablation results showed that fold-specific standardisation was essential for RBF-SVM and that increasing convolutional-kernel size did not consistently improve performance. The study is limited by its use of a preprocessed binary benchmark dataset and the absence of verified patient identifiers. Future work should evaluate the framework using subject-wise and cross-dataset validation, larger clinical EEG datasets, multi-class seizure states, and more extensive interpretability analysis.

All the Declarations and Statements

Author Contributions Statement
Mrs. Sanagavarapu Sunitha contributed to conceptualization, methodology development, software implementation, data processing, formal analysis, validation, visualization, and preparation of the original manuscript.
Dr. Umadevi Ramamoorthy contributed to conceptualization, methodological supervision, validation, interpretation of the findings, critical review, editing, and overall research supervision. Both authors reviewed and approved the final manuscript.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement
The authors declare that there are no financial, professional, or personal conflicts of interest that could have influenced the research reported in this manuscript.

Funding Declaration
No external funding was received for this research.

Data Availability Statement
The Epileptic Seizure Recognition dataset used in this study is publicly available on Kaggle at:

https://www.kaggle.com/datasets/harunshimanto/epileptic-seizure-recognition

The processed data tables, recording-group assignments, outer- and inner-fold partitions, prediction files, and derived experimental results are available from the corresponding author upon reasonable request.

Ethical Declarations
Not applicable. The study used a publicly available, preprocessed, and de-identified benchmark dataset. No new data were collected from human participants or animals, and no direct interaction with patients was conducted.

Acknowledgments
The authors acknowledge Harun-Ur-Rashid for making the Epileptic Seizure Recognition dataset publicly available on Kaggle. The dataset was derived from the EEG recordings originally introduced by Andrzejak et al. and has been widely used as a benchmark for epileptic seizure-detection research.

Declaration of Generative AI in Scholarly Writing
N/A.

Abbreviations
The following abbreviations are used in this manuscript:

EEG - Electroencephalogram
AI - Artificial Intelligence
ML – Machine Learning
DL - Deep Learning
SVM – Support Vector Machine
RF – Random Forest
kNN – k-Nearest Neighbors
RBF – Radial Basis Function
MLP – Multi-Layer Perceptron
CNN – Convolutional Neural Networks
CNN 1D – 1-Dimensional Convolutional Neural Networks
EEGNet 1D – 1 Dimensional Electroencephalogram Neural Network
RNN - Recurrent Neural Networks
TCN – Temporal Convolutional Network
LSTM – Long Short Term Memory
ROCKET – RandOm Convolutional Kernel Transform
MultiROCKET – Multi-resolution RandOm Convolutional Kernel Transform
MiniROCKET – Minimal RandOm Convolutional Kernel Transform

AUC – Area Under the Curve
ROC AUC – Receiver Operating Characteristic – Area Under the Curve

Appendix A/B/C…, with appendix title

N/A.

References

[1] Hosseinzadeh, M., P. Khoshvaght, S. Sadeghi, P. Asghari, A. N. Varzeghani, M. Mohammadi, et al., (2024) "A Model for Epileptic Seizure Diagnosis Using the Combination of Ensemble Learning and Deep Learning," IEEE Access.

[2] Nhu, D., M. Janmohamed, L. Shakhatreh, O. Gonen, P. Perucca, A. Gilligan, et al., (2023) "Automated Interictal Epileptiform Discharge Detection from Scalp EEG Using Scalable Time-Series Classification Approaches," International Journal of Neural Systems 33(1), 2350001.

[3] Lytton, W. W., (2008) "Computer Modelling of Epilepsy," Nature Reviews Neuroscience 9(8), 626–637.

[4] Zamanzadeh Darban, Z., G. I. Webb, S. Pan, C. Aggarwal, and M. Salehi, (2024) "Deep Learning for Time Series Anomaly Detection: A Survey," ACM Computing Surveys 57(1), 1–42.

[5] Siripurapu, S., N. K. Darimireddy, A. Chehri, B. Sridhar, and A. V. Paramkusam, (2023) "Technological Advancements and Elucidation Gadgets for Healthcare Applications: An Exhaustive Methodological Review—Part I (AI, Big Data, Blockchain, Open-Source Technologies, and Cloud Computing)," Electronics 12(3), 750.

[6] Harun-Ur-Rashid. (2017). Epileptic Seizure Recognition Dataset. Kaggle. https://www.kaggle.com/datasets/harunshimanto/epileptic-seizure-recognition.

[7] Magnuson, J., F. Leonessa, and G. S. Ling, (2012) "Neuropathology of Explosive Blast Traumatic Brain Injury," Current Neurology and Neuroscience Reports 12(5), 570–579.

[8] Kode, H., K. Elleithy, and L. Almazaydeh, (2024) "Epileptic Seizure Detection in EEG Signals Using Machine Learning and Deep Learning Techniques," IEEE Access 12, 80657–80668.

[9] Mohammadi Foumani, N., L. Miller, C. W. Tan, G. I. Webb, G. Forestier, and M. Salehi, (2024) "Deep Learning for Time Series Classification and Extrinsic Regression: A Current Survey," ACM Computing Surveys 56(9), 1–45.

[10] Liu, R., S. Wang, F. Tian, and L. Yi, (2025) "SIR-3DCNN: A Framework of Multivariate Time Series Classification for Lung Cancer Detection," IEEE Transactions on Instrumentation and Measurement.

[11] Jahan, S., F. Nowsheen, M. M. Antik, M. S. Rahman, M. S. Kaiser, A. S. Hosen, and I. H. Ra, (2023) "AI-Based Epileptic Seizure Detection and Prediction in Internet of Healthcare Things: A Systematic Review," IEEE Access 11, 30690–30725.

[12] Chen, R., X. Yan, G. Xiao, Y. Hou, and S. Zhou, (2025) "Trend and Order Features for Semi-Supervised Time-Series Classification via Multitask Learning," IEEE Transactions on Neural Networks and Learning Systems.

[13] Tzallas, A. T., M. G. Tsipouras, D. G. Tsalikakis, E. C. Karvounis, L. Astrakas, S. Konitsiotis, and M. Tzaphlidou, (2012) "Automated Epileptic Seizure Detection Methods: A Review Study," Epilepsy—Histological, Electroencephalographic and Psychological Aspects, 75–98.

[14] Shoeibi, A., M. Khodatars, N. Ghassemi, M. Jafari, P. Moridian, R. Alizadehsani, et al., (2021) "Epileptic Seizures Detection Using Deep Learning Techniques: A Review," International Journal of Environmental Research and Public Health 18(11), 5780.

[15] Siddiqui, M. K., R. Morales-Menendez, X. Huang, and N. Hussain, (2020) "A Review of Epileptic Seizure Detection Using Machine Learning Classifiers," Brain Informatics 7(1), 5.

[16] Vidyaratne, L. S., and K. M. Iftekharuddin, (2017) "Real-Time Epileptic Seizure Detection Using EEG," IEEE Transactions on Neural Systems and Rehabilitation Engineering 25(11), 2146–2156.

[17] Tran, L. V., H. M. Tran, T. M. Le, T. T. Huynh, H. T. Tran, and S. V. Dao, (2022) "Application of Machine Learning in Epileptic Seizure Detection," Diagnostics 12(11), 2879.

[18] Yu, S., R. El Atrache, J. Tang, M. Jackson, A. Makarucha, S. Cantley, et al., (2023) "Artificial Intelligence-Enhanced Epileptic Seizure Detection by Wearables," Epilepsia 64(12), 3213–3226.

[19] Paul, Y., (2018) "Various Epileptic Seizure Detection Techniques Using Biomedical Signals: A Review," Brain Informatics 5(2), 6.

[20] Sadati, N., H. R. Mohseni, and A. Maghsoudi, (2006) "Epileptic Seizure Detection Using Neural Fuzzy Networks," Proceedings of the 2006 IEEE International Conference on Fuzzy Systems, Vancouver, Canada, 596–600.

[21] Guo, L., D. Rivero, J. Dorado, J. R. Rabuñal, and A. Pazos, (2010) "Automatic Epileptic Seizure Detection in EEGs Based on Line Length Feature and Artificial Neural Networks," Journal of Neuroscience Methods 191(1), 101–109.

[22] Andrzejak, R. G., Lehnertz, K., Rieke, C., Mormann, F., David, P., and Elger, C. E. (2001). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. Physical Review E, 64(6), 061907. https://doi.org/10.1103/PhysRevE.64.061907.

[23] Lawhern, V. J., A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, (2018) "EEGNet: A Compact Convolutional Neural Network for EEG-Based Brain–Computer Interfaces," *Journal of Neural Engineering* 15(5), 056013.

[24] Dempster, A., F. Petitjean, and G. I. Webb, (2020) "ROCKET: Exceptionally Fast and Accurate Time Series Classification Using Random Convolutional Kernels," *Data Mining and Knowledge Discovery* 34(5), 1454–1495.

[25] Dempster, A., D. F. Schmidt, and G. I. Webb, (2021) "MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification," *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 248–257.

[26] Tan, C. W., A. Dempster, C. Bergmeir, and G. I. Webb, (2022) "MultiRocket: Multiple Pooling Operators and Transformations for Fast and Effective Time Series Classification," *Data Mining and Knowledge Discovery* 36(5), 1623–1646.

Authors' Profiles



Sanagavarapu Sunitha is pursuing her Ph.D. in Computer Science at CMR University, Bengaluru, India. She earned her M.Sc. in Computer Science from CMR University and her B.Sc. in Computer Science from Andhra University, India. She is currently an interdisciplinary academic researcher at CMR University, specializing in healthcare data science, with a particular focus on EEG-based seizure detection using machine learning. Her research interests include artificial intelligence for sustainable power systems, digital health, ethical AI deployment, and advanced machine-learning techniques for biomedical signal analysis. She is actively involved in academic publishing, conference participation, and professional networking through IEEE.



Dr. Umadevi Ramamoorthy is an Associate Professor in the School of Science and Computer Studies at CMR University, Bengaluru, Karnataka, India. She holds a Ph.D. in Computer Science and has 20 years of teaching experience and 8 years of research experience. She has presented and published research papers in several international and national conferences and in journals indexed in SCI Expanded, Scopus, UGC-CARE, and the UGC-approved list. She is a member of IAENG and AACST. She has published patents and received a design patent grant from the Intellectual Property Office, Government of India. She currently supervises six research scholars and has received several recognitions, including the Best Young Scientist Award, Best Academician Award, Best Paper Award, and Research Excellence Award. She serves on the editorial boards of several journals and acts as a reviewer for reputed national and international journals. She has also been invited as a resource person for various academic events at the national and international levels. Her areas of expertise include secure communication, information security, and artificial intelligence.

How to cite this paper

Sanagavarapu Sunitha, Umadevi Ramamoorthy, "Leakage-Controlled Recording-Level Evaluation of Machine Learning, Deep Learning, and Convolutional-Kernel Models for EEG-Based Epileptic Seizure Detection", International Journal of Engineering and Manufacturing (IJEM), Vol.16, No.5, pp. 179-204, 2026. DOI:10.5815/ijem.2026.05.10