POLITECNICO DI TORINO
Repository ISTITUZIONALE

An Integrated Calibration–Uncertainty Framework for Improving the Reliability of Deep Learning Models for Seizure Detection in EEG Signals

Original
An Integrated Calibration–Uncertainty Framework for Improving the Reliability of Deep Learning Models for Seizure Detection in EEG Signals / Seoni, S., Molinari, F., Beneveri, M., Salvi, M.. - In: COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE. - ISSN 0169-2607. - 285:(2026). [10.1016/j.cmpb.2026.109520]

Availability:
This version is available at: 11583/3012167 since: 2026-06-17T20:03:20Z

Publisher:
Elsevier

Published
DOI:10.1016/j.cmpb.2026.109520

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

20 November 2026

Computer Methods and Programs in Biomedicine 285 (2026) 109520



Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine





An integrated calibration–uncertainty framework for improving the reliability of deep learning models for seizure detection in EEG signals

Silvia Seoni *, Filippo Molinari, Margherita Benevieri, Massimo Salvi

Biolab, PolitoBIOMedLab, Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy

ARTICLE INFO

Keywords:
Calibration
EEG
Monte Carlo dropout
Seizure detection
Uncertainty quantification

ABSTRACT

Background and objective: Deep learning models have demonstrated strong performance in automated seizure detection from EEG signals. However, these models may produce confident predictions even when incorrect, limiting their reliability barrier for clinical adoption. This study proposes an integrated calibration–uncertainty framework to enhance model reliability in EEG-based seizure classification.

Methods: A CNN–BiLSTM model was trained to classify EEG epochs containing epileptic seizure activity. The framework leverages Expected Calibration Error (ECE) to assess global confidence reliability and Monte Carlo Dropout (MCD)-based uncertainty quantification to identify unreliable predictions. For each dropout rate, we evaluated both model calibration and the entropy-based separability between correctly (CC) and misclassified (MC) samples, computed as the Overlap Area between their uncertainty distributions. A multi-objective selection strategy was then used to automatically identify the configuration that best balances these complementary aspects. Finally, a selective classification approach was implemented, using an uncertainty threshold to identify unreliable predictions and defer them for further clinical evaluation.

Results: Varying the dropout rate significantly affected both calibration and uncertainty behaviour. The optimal balance was achieved at $p = 0.1$, yielding the lowest combined ECE and Overlap Area. The selective classification improved accuracy from 91.7% (baseline) to 99.6% while retaining ~74% of samples, outperforming models optimized for either calibration or uncertainty alone.

Conclusions: The proposed dual perspective framework improves model robustness by integrating global confidence calibration with local uncertainty estimation, representing a practical step toward reliable AI deployment in clinical neurophysiology.

1. Introduction

Epilepsy is a chronic neurological disorder affecting approximately 1% of the global population and is characterized by recurrent, unprovoked seizures [1]. Electroencephalography (EEG) remains the gold standard for detecting and characterizing epileptic activity, providing a non-invasive tool for both clinical diagnosis and continuous monitoring [2].

In recent years, artificial intelligence (AI) and, in particular, deep learning (DL) models have greatly advanced EEG-based seizure detection, achieving near-human accuracy in distinguishing seizure (ictal) and non-seizure (interictal) events [2,3]. Among these, one-dimensional convolutional neural networks (1D-CNNs) have been proposed to operate directly on raw multi-channel EEG signals, automatically learning local temporal filters and discriminative waveform patterns [4,5]. Conversely, two-dimensional CNNs (2D-CNNs) exploit time–frequency transformations, such as wavelet or spectrogram representations, to capture both spectral and spatial characteristics of seizure activity [6]. Recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM) units, effectively model the temporal dynamics of seizure evolution by exploiting long-term dependencies within the EEG sequence [7]. Hybrid CNN–BiLSTM models further improve performance by combining the spatial sensitivity of CNNs with the temporal modeling capabilities of recurrent layers [8].

Despite their impressive performance, DL models are not yet widely adopted in clinical practice. This limited translation stems from reliability issues in their predictions. While these models provide

* Corresponding author at: Biolab, Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy, Corso Duca degli Abruzzi, 24, 10129 Turin, Italy.
E-mail address: silvia.seoni@polito.it (S. Seoni).

https://doi.org/10.1016/j.cmpb.2026.109520
Received 4 February 2026; Received in revised form 10 June 2026; Accepted 11 June 2026
Available online 12 June 2026
0169-2607/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

confidence scores (typically through softmax probabilities) along with their predictions, these scores can be misleadingly high even for incorrect predictions, undermining their practical reliability. Indeed, AI models are often poorly calibrated and tend to assign excessively high confidence to incorrect predictions, which compromises their reliability in clinical decision-making [9].

To address this limitation, two main approaches have emerged: calibration methods and uncertainty quantification (UQ). Calibration methods aim to ensure that predicted probabilities align with true correctness likelihood at a population level, with Expected Calibration Error (ECE) being one of the most widely adopted metrics [10]. Conversely, UQ focuses on individual predictions, providing tools to estimate the trustworthiness of each specific output [11].

Recent studies have applied calibration methods, particularly ECE, to evaluate the reliability of DL models in medical imaging [9], but applications to EEG analysis remain limited. While this approach can improve global model reliability [10], it cannot address the need for sample-specific reliability assessment, which is crucial in clinical applications where individual predictions can impact patient care.

UQ complements this global perspective by providing insight into the trustworthiness of individual predictions [11], which is crucial in safety-critical contexts such as seizure detection [12]. However, most existing studies treat UQ as a post-hoc analysis, using it primarily for selective classification [13]. For example, Wong et al. [14] proposed a novel ensemble model to identify low-confidence predictions in seizure detection, achieving 87% accuracy. While effective, this approach does not exploit the synergy between uncertainty information and model calibration.

These two perspectives (i.e., global calibration and individual uncertainty) represent complementary aspects of model reliability. Their integration within a unified framework offers an opportunity to develop models that are both globally calibrated and locally uncertainty-aware, thereby enhancing the robustness, interpretability, and clinical usability of AI-based EEG analysis.

Jiahao et al. [15] integrated Bayesian inference with post-hoc calibration for seizure detection, achieving a 38% reduction in Negative Log Likelihood and a 43% reduction in Brier score. In non-clinical settings, studies have explored how acquisition-related uncertainty and model confidence jointly affect the robustness of EEG architectures under realistic data perturbations [16]. However, a systematic framework for leveraging both calibration and uncertainty in clinical EEG analysis remains unexplored.

In this work, we propose an integrated calibration–uncertainty framework that jointly analyzes model calibration (ECE) and predictive uncertainty (entropy-based separability) across different dropout rates. This approach quantifies how stochasticity affects confidence alignment and uncertainty discrimination, enabling the automatic identification of an optimal operating configuration. Finally, a selective classification mechanism based on predictive uncertainty is implemented, allowing the model to abstain from unreliable predictions and thus improve practical reliability. The key methodological and experimental contributions of this study can be summarized as follows:

- Development of an integrated calibration–uncertainty framework to enhance the reliability of deep learning models for EEG-based seizure detection.
- Systematic analysis of the stochastic effect of dropout on both confidence calibration and predictive uncertainty, revealing their complementary relationship.
- An automated selection strategy that combines calibration (ECE) and uncertainty (Overlap Area) metrics to objectively determine the optimal dropout configuration for Monte Carlo Dropout (MCD) inference.
- Implementation and validation of a selective classification approach that enhances effective accuracy and supports safe deployment of AI-assisted EEG diagnostic systems.

2. Materials and methods

2.1. Dataset

The experiments were conducted using the Children's Hospital Boston–MIT (CHB–MIT) Scalp EEG Database, one of the most established and publicly available datasets for epilepsy research, particularly for pediatric patients [17,18]. The dataset consists of recordings from 24 subjects aged between 1.5 and 22 years, both male and female, all diagnosed with intractable epilepsy. EEG signals were acquired using 21 scalp electrodes arranged according to the international 10–20 system, with a sampling frequency of 256 Hz. Each recording session lasted between one and four hours, including both ictal and interictal periods, for a total of over one hundred hours of EEG activity. Seizure onsets and durations were annotated by clinical experts, totalling 198 seizure events.

For this study, a subset of 17 pediatric patients was selected to ensure homogeneous electrode montages and consistent data quality. Patients with irregular channel configurations (e.g., Chb12, Chb13) or with recordings not representative of the pediatric cohort (e.g., ages above 16 years or seizures shorter than 10 s) were excluded [19]. The final cohort included 103 seizures, with a total ictal duration of approximately 6684 s. This selection strategy followed previous CHB–MIT studies and was adopted to improve inter-patient consistency while reducing variability related to heterogeneous electrod montages.

All EEG signals were processed using the MNE-Python framework to harmonize channel configurations. Auxiliary and duplicate channels (e.g., ECG, VNS, or mislabeled electrodes) were removed, resulting in a canonical set of 21 channels for all recordings. Signals were filtered using a finite impulse response (FIR) band-pass filter with cut-off frequencies between 0.01 and 128 Hz, preserving clinically relevant EEG patterns while attenuating low-frequency drift and high-frequency noise.

The continuous EEG recordings were segmented into non-overlapping 4-second windows, corresponding to 1024 samples per channel. This window length has been reported as an optimal compromise between temporal resolution and contextual information for seizure detection in deep learning models [19]. Segments were labeled as seizure (ictal) if fully contained within annotated seizure intervals and as non-seizure (interictal) if located at least 30 s away from any seizure event. This margin was adopted as a conservative criterion to reduce ambiguity near seizure boundaries, where the transition between ictal and interictal states may be gradual and prone to label noise. To mitigate class imbalance, an equal number of interictal segments was randomly selected for each patient, matching the number of ictal segments. Each segment was normalized using per-channel z-score normalization, where the mean and standard deviation were computed across all segments and time samples within each dataset split (training, validation, and test), ensuring zero mean and unit variance while avoiding information leakage.

The dataset was divided on a patient-wise basis into training (80%), validation (10%), and test (10%) subsets to evaluate the generalization capability of the models to unseen subjects, a critical aspect for clinical reliability. The final partition included 13 subjects for training, 2 for validation, and 2 for testing, as summarized in Table 1.

Table 1
The number of epochs for training, validation, and test set.

Set	Total segments (Ictal)	Subjects
Training	2594 (1297)	chb02, chb03, chb05, chb06, chb07, chb08, chb09, chb11, chb14, chb17, chb20, chb23, chb24
Validation	318 (159)	chb10, chb22
Test	314(157)	chb01, chb21

2.2. Model architecture

A hybrid convolutional–recurrent neural network (CNN–BiLSTM) was implemented to classify EEG epochs into seizure (ictal) and non-seizure (interictal) events.

The architecture combines convolutional layers for spatial–spectral feature extraction with bidirectional LSTM layers for temporal pattern modeling. The final network includes a temporal convolutional front-end (six 1D convolutional layers with batch normalization and max pooling), a two-layer BiLSTM module, and a fully connected classification block with softmax output over two classes.

Three dropout layers (p = 0.1) were included to regularize training and were later used for MCD inference to estimate predictive uncertainty [20]. The network was implemented in *TensorFlow 2.0* with *Keras 3.10.0* as a backend.

The model was trained end-to-end in a supervised fashion using the Adam optimizer (learning rate = 1×10^{-4}) and sparse categorical cross-entropy loss. Training was performed with a batch size of 32 samples for up to 100 epochs, applying early stopping based on validation accuracy with a patience of 5 epochs to avoid overfitting. Dropout was active during training and reactivated at inference time for uncertainty estimation. A summary of the implemented CNN–BiLSTM architecture is reported in Table 2.

2.3. Uncertainty quantification

Uncertainty estimation was performed using the MCD approach [20]. This method allows quantifying predictive uncertainty by assessing the variability of the model's probabilistic outputs across multiple forward passes. For each dropout probability ($p \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$), N = 10 stochastic forward passes were performed on the validation set while keeping dropout active. For each sample, the normalized predictive entropy was computed as the measure of uncertainty according to [21]:

$$H_{norm}(x) = -\frac{1}{\log C}\sum_{i=1}^{C} p_i(x)\log p_i(x) \tag{1}$$

where C is the number of classes (in this case, $C = 2$) and $p_i(x)$ is the softmax probability assigned to class i. Low entropy values indicate confident predictions, while high entropy reflects uncertain or inconsistent model outputs.

Table 2
CNN-BiLSTM architecture.

Block	Layers	Output Channels/ Units	Dropout
Input	EEG epoch (1024 × 21)	–	–
Block 1	Conv1D + BN + ReLU → MP1D	8	–
Block 2	Conv1D + BN + ReLU → MP1D	16	-
Block 3	Conv1D + BN + ReLU → MP1D → **Dropout**	24	0.1
Block 4	Conv1D + BN + ReLU → MP1D	36	-
Block 5	Conv1D + BN + ReLU → MP1D	48	–
Block 6	Conv1D + BN + ReLU → MP1D → **Dropout**	56	0.1
Recurrent module	2 × BiLSTM → **Dropout**	[64, 32]	0.1
Classifier	Dense (50) → Dense (20) → Softmax (2)	2 (classes)	-

Conv1D denotes a one-dimensional convolutional layer, BN stands for Batch Normalization, ReLU refers to the Rectified Linear Unit activation function, and MP1D indicates MaxPool1D, a one-dimensional max pooling layer. Linear corresponds to a fully connected (dense) layer.

The entropy values were then grouped into two categories:

- Correctly classified (CC) samples, which are the predictions matching the true label,
- Misclassified (MC) samples, which are the incorrect predictions.

For each dropout rate p, we analyzed how effectively model uncertainty (entropy) distinguished between these two groups. In an effective uncertainty-aware model, misclassified samples are expected to exhibit higher uncertainty than correctly classified ones. To quantify this separation, we:

1. Estimated the distribution of entropy values for both CC and MC samples using kernel density estimation.
2. Computed the Overlap Area (OVL) between these two distributions as:

$$\mathrm{OVL} = \int \min\left(f_{\mathrm{CC}}(h), f_{\mathrm{MC}}(h)\right) dh \tag{2}$$

where $f_{\mathrm{CC}}(h)$and $f_{\mathrm{MC}}(h)$ denote the normalized entropy density functions of the CC and MC samples, respectively. Smaller values of OVL indicate better separation between the CC and MC distributions, meaning the model's uncertainty more effectively distinguishes between reliable and unreliable predictions. For example, an OVL of 0 would indicate perfect separation, while an OVL of 1 would indicate complete overlap.

By computing the OVL across different dropout rates, we identified the optimal rate p that maximizes the discriminative power of uncertainty estimation. This optimal value was then used for subsequent calibration analysis and selective classification experiments.

2.4. Calibration model

To assess the model's confidence and evaluate the effect of stochasticity introduced by dropout, a calibration analysis was performed across multiple dropout rates ($p \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$) in the validation set. For each configuration, the ECE was computed, and the corresponding reliability diagram (accuracy vs. confidence plot) was generated to visualize the relationship between predicted confidence and empirical accuracy [9,22]. In addition, the ECE and reliability diagram were also calculated for the baseline model (without MCD) to provide a reference comparison.

The ECE provides a quantitative measure of miscalibration by partitioning the predicted confidence values into M equally spaced bins and calculating the weighted average of the absolute difference between accuracy and confidence within each bin:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{3}$$

where B_m denotes the set of samples whose confidence scores fall into the m-th bin, n is the total number of samples, $acc(B_m)$ is the empirical accuracy, and $conf(B_m)$ is the mean predicted confidence within that bin. Lower ECE values indicate better calibration, meaning that the model's predicted probabilities align more closely with the actual likelihood of correctness.

The dropout configuration associated with the lowest ECE was considered the best-calibrated setting and was used as a reference for comparison with the uncertainty-based and integrated selection criteria.

2.5. Integrated calibration–uncertainty selection strategy

Optimizing calibration and uncertainty separately leads to different optimal dropout rates, as these aspects capture different reliability

characteristics. Our integrated approach aims to find a balanced configuration that achieves both good calibration and effective separation between correct and incorrect predictions.

For each dropout probability, the ECE and the OVL between predictive-uncertainty distributions were computed on the validation set. After normalizing these metrics using min–max scaling, we selected the optimal dropout configuration by minimizing the weighted Euclidean distance from the ideal point (0,0), representing perfect calibration and maximal uncertainty separability:

$$D_w(p) = \sqrt{(w_{ECE}\widetilde{ECE}(p))^2 + (w_{OVL}\widetilde{OVL}(p))^2}. \quad (4)$$

The weighting factors were set to $w_{\mathrm{ECE}} = 0.4$ and $w_{\mathrm{OVL}} = 0.6$, giving slightly higher relevance to uncertainty separability for clinical reliability. This criterion identified $p = 0.1$ as the optimal configuration, yielding the best quantitative trade-off between calibration and uncertainty discrimination under MCD inference.

2.6. Selective classification

To further enhance model reliability, we implemented a selective classification framework where the model can abstain from predictions considered unreliable based on their predictive uncertainty (Fig. 1). This approach introduces a trade-off between accuracy and data coverage. In this setting, samples associated with high uncertainty are not removed from the analysis, but are instead identified as requiring further clinical evaluation. This allows the model to distinguish between reliable predictions and uncertain cases, supporting the clinician in focusing on segments that may benefit from additional inspection, while preserving all potentially relevant information.

The discriminative uncertainty threshold (τ) was initially estimated as the intersection point between the entropy distributions of CC and MC samples obtained from the validation set. Predictions with uncertainty values above τ were considered unreliable and were deferred for further clinical evaluation, while those below τ were retained as "confident" samples.

The optimal uncertainty threshold τ^* was selected on the validation set by maximizing a utility function that balances model accuracy and data retention:

$$U_\gamma = \mathrm{Accuracy} \cdot \left(\frac{\mathrm{Coverage}}{100}\right)^\gamma \quad (5)$$

where $\gamma = 0.10$ provides a mild emphasis on accuracy while preserving sufficient coverage. This criterion ensures that the selected threshold maximizes the practical reliability of the model by jointly considering predictive performance and the fraction of retained samples.

Once the optimal threshold τ^* was determined, it was applied to the test set to evaluate generalization performance. Selective classification was implemented for the three model configurations obtained from the different *p*-selection strategies:

1. ECE-only: model with minimum calibration error,
2. Uncertainty-only: model with minimum OVL,
3. Integrated: model balancing both calibration and uncertainty (trade-off criterion).

After applying the optimal threshold, all performance metrics - coverage, accuracy, specificity, sensitivity, and F1-score - were recomputed to assess how each selection strategy affected model reliability and data retention.

3. Results

3.1. Baseline model performance

Before integrating calibration and uncertainty quantification, the CNN–BiLSTM model was evaluated in its deterministic configuration (dropout disabled during inference, $p = 0$) to establish a baseline reference.

On the test set, the model achieved an overall accuracy of 0.917, confirming a solid classification capability in distinguishing seizure (ictal) from non-seizure (interictal) EEG segments. For the interictal class (class 0), the model obtained a precision of 0.88, a sensitivity of 0.97, and an F1-score of 0.92. For the ictal class (class 1), the corresponding metrics were precision of 0.96, sensitivity of 0.87, and F1-score of 0.91, indicating balanced detection performance across both categories.

3.2. Effect of dropout on predictive uncertainty

Fig. 2 shows, for each dropout rate (*p*), the entropy distributions of CC and MC samples (bottom panel) and the corresponding OVL values (top panel). At low dropout probabilities, the two distributions appear more clearly separated, with MC samples exhibiting markedly higher entropy values than CC samples. As *p* increases, the two distributions progressively converge, indicating reduced discriminability between

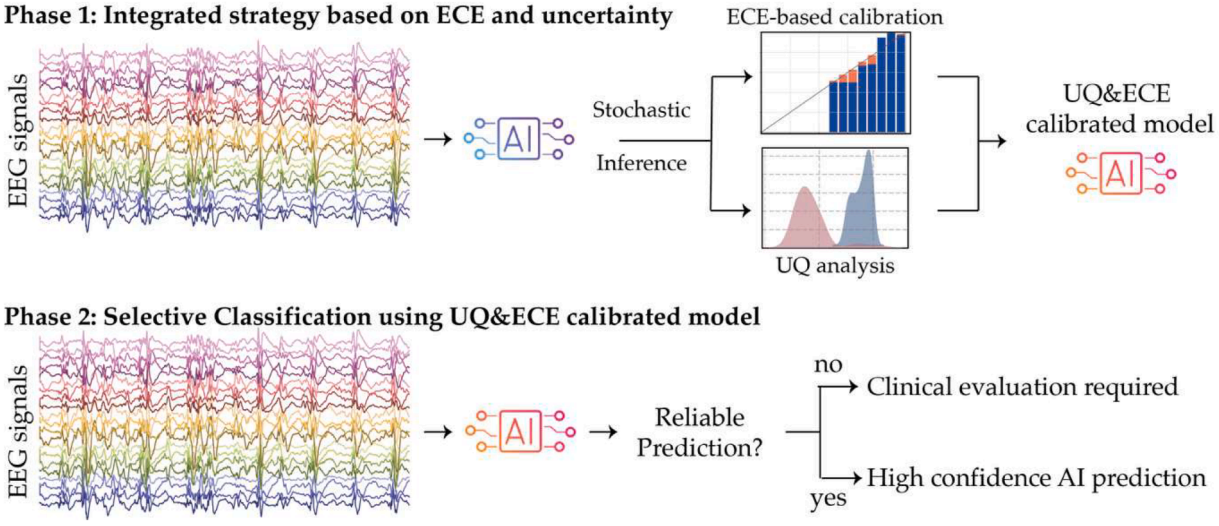


Fig. 1. Overview of the proposed pipeline. Phase 1 performs integrated calibration using uncertainty estimation and ECE analysis. Phase 2 applies the calibrated model to seizure detection, distinguishing between reliable predictions and uncertain samples that require further clinical evaluation.

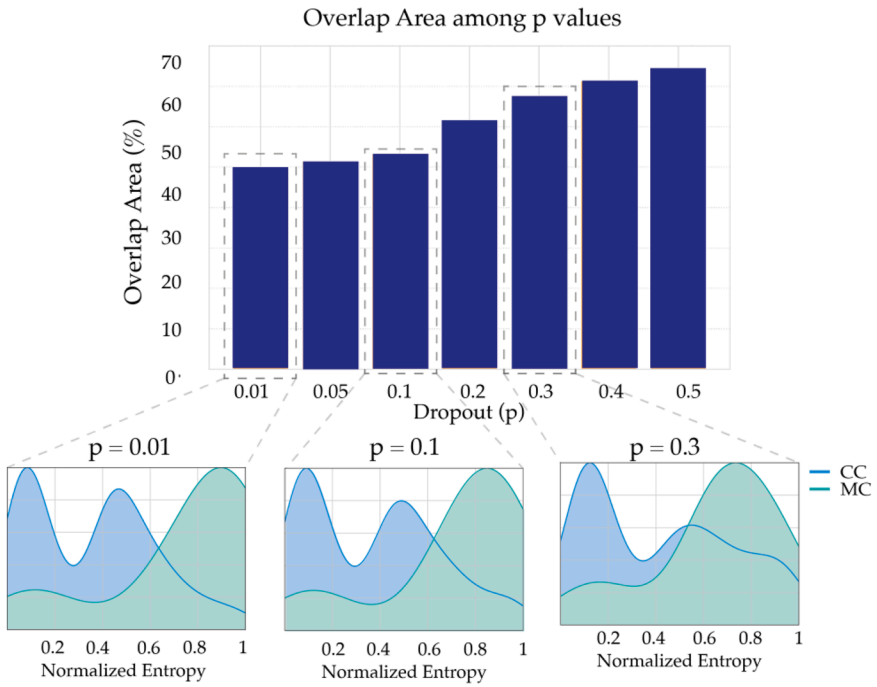


Fig. 2. Predictive uncertainty distributions and Overlap Area (OVL) across dropout rates. (Top) Bar plot of the OVL values as a function of dropout rate (p). (Bottom) Normalized entropy distributions for correctly classified (CC, blue) and misclassified (MC, green) samples.

correct and incorrect predictions. The OVL quantitatively captures this effect, showing lower values for smaller dropout rates ($p = 0.01$, 0.05, 0.1). A smaller OVL value indicates higher separability and, consequently, a stronger ability of the model's uncertainty to distinguish reliable from unreliable predictions.

As reported in Fig. 2, the OVL decreases for small dropout rates, reaching its minimum at $p = 0.01$, and then increases again for higher values. Based on this analysis, $p = 0.01$ was selected as the uncertainty-optimal configuration, corresponding to the dropout rate that provides the best discriminative power between reliable and unreliable samples.

3.3. Effect of dropout on model calibration

Fig. 3 presents the reliability diagrams for all dropout rates and the corresponding ECE values. A well-calibrated model should exhibit points lying close to the diagonal, where predicted confidence matches empirical accuracy. At low dropout probabilities, the model tends to be slightly overconfident, while higher dropout values introduce underconfidence and increase variability in the predicted scores.

The lowest ECE was observed at $p = 0.3$ (ECE = 0.042), indicating the best alignment between the confidence and accuracy. Based on this analysis, $p = 0.3$ was selected as the calibration-optimal configuration, representing the dropout rate providing the most reliable confidence estimates.

3.4. Integrated model selection

Fig. 4 shows the behavior of the weighted distance $D_w(p)$ across dropout rates. The minimum distance occurred at $p = 0.1$, corresponding to the configuration that provides the best quantitative trade-off between calibration (ECE = 0.061) and uncertainty discrimination (OVL = 0.53). This dropout rate was therefore selected as the integrated

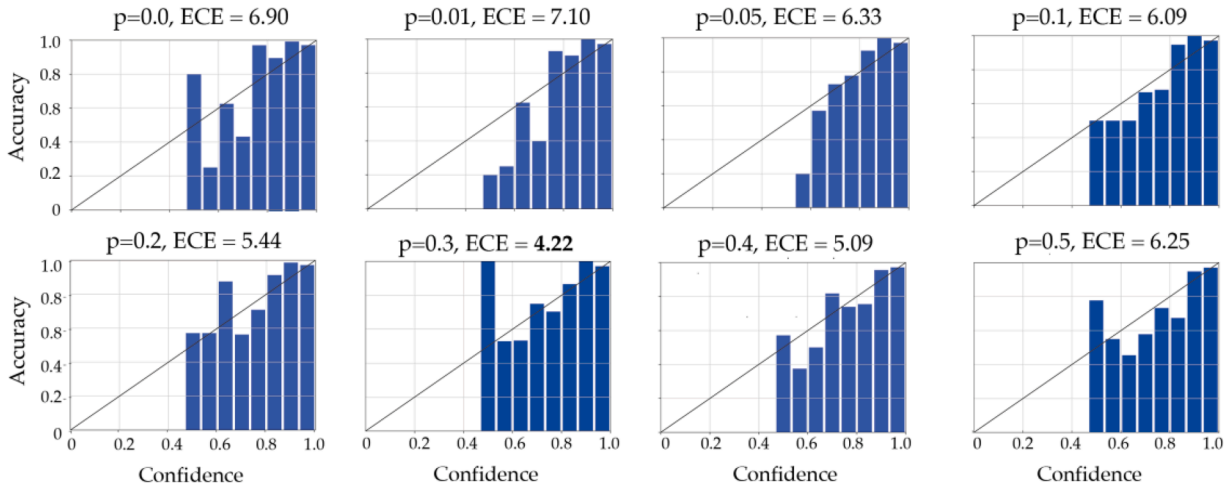


Fig. 3. Model calibration across dropout rates. Reliability diagrams showing the empirical accuracy for different dropout probabilities and the corresponding ECE values for each configuration.

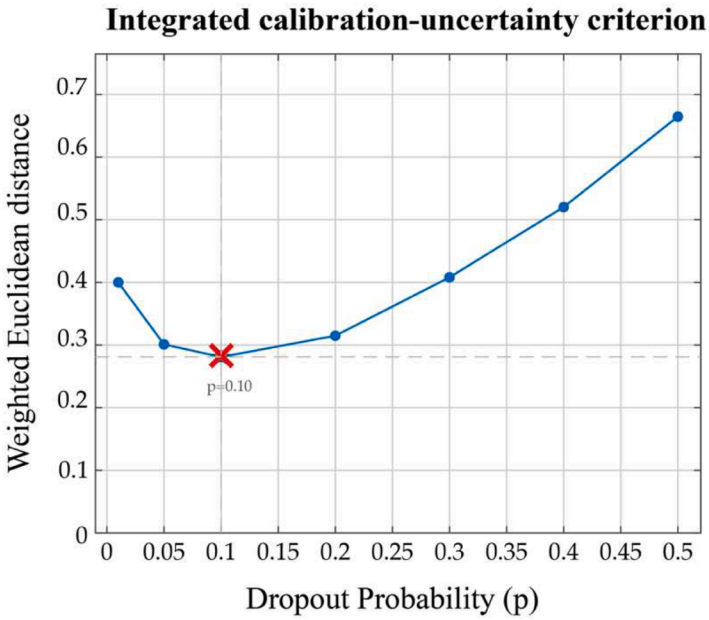


Fig. 4. Weighted distance-based integrated model selection.

configuration for the subsequent selective-classification analysis. Neighboring configurations ($p = 0.05$ and $p = 0.2$) yielded comparable distance values, confirming that the selected point lies within a stable region of the reliability curve.

3.5. Selective classification

The final step of the proposed framework aimed to evaluate how uncertainty-based filtering can enhance model reliability through selective classification. The optimal uncertainty thresholds (τ^*) for each configuration were determined on the validation set by maximizing the utility function defined in Eq. (5), which balances model accuracy and data retention.

The resulting optimal thresholds were $\tau^* \approx 0.66$ for $p = 0.01$ (UQ-only), $\tau^* \approx 0.68$ for $p = 0.3$ (ECE-only), and $\tau^* \approx 0.72$ for $p = 0.1$ (UQ&ECE). These thresholds were then applied to the test set to assess the generalization of the selective classification strategy.

Table 3 reports the classification performance and highlights how uncertainty-aware configurations ($p = 0.01$ and $p = 0.1$) lead to more effective filtering strategies, achieving higher accuracy while retaining a larger portion of samples. In particular, the uncertainty-based configuration ($p = 0.01$) reached an accuracy of 0.996 with a coverage of 72% at an optimal uncertainty threshold $\tau^* = 0.66$. The integrated configuration ($p = 0.1$) achieved comparable accuracy (0.996) with slightly higher coverage (74%), indicating a more selective yet efficient filtering strategy. Conversely, the calibration-based model ($p = 0.3$) achieved an accuracy of 0.989 but with a lower coverage (69%), confirming that excessive focus on calibration can reduce data retention without substantial accuracy gains.

A noticeable difference was also observed in specificity values, which were higher for the uncertainty-based and integrated filtering strategies (0.99) and lower for the calibration-only model. This indicates that the uncertainty-driven approaches are less prone to generating false positives, further improving the model's reliability in a clinical setting. The integrated configuration ultimately provided the best trade-off between performance and coverage, retaining 2% more samples than the uncertainty-only model while maintaining the same accuracy.

Fig. 5 reports the confusion matrices for the baseline model and for the three proposed filtering strategies. Across all methods, filtering effectively removed false-negative samples, which represent the most critical errors in a clinical context, as they correspond to missed seizure events. The strategy based solely on calibration (ECE filtering, $p = 0.3$) further reduced false negatives but at the cost of a larger number of discarded true negatives, resulting in a lower overall coverage.

Conversely, the uncertainty-based (UQ filtering, $p = 0.01$) and integrated (UQ & ECE filtering, $p = 0.1$) strategies maintained a higher proportion of correctly classified non-seizure samples while still eliminating unreliable predictions.

Notably, the integrated approach achieved the same low false positive rate as the UQ-only strategy (just 1 false positive) while preserving more informative samples, specifically retaining 2 additional true negatives and 3 additional true positives compared to using uncertainty alone. This demonstrates that considering both calibration and uncertainty enables more selective and efficient filtering, leading to the most balanced outcome that combines high classification accuracy with optimal sample retention.

Table 3
Performance comparison between the proposed filtering strategies and baseline model.

Model	Acc. (Cov.)	Sens.	Spec.	F1 score
Baseline (no filtering)	0.917 (100%)	0.87	0.96	0.91
UQ-only ($p = 0.01$)	0.996 (72%)	1.00	0.99	0.99
ECE-only ($p = 0.3$)	0.989 (69%)	1.00	0.89	0.99
UQ&ECE ($p = 0.1$)	0.996 (74%)	1.00	0.99	0.99

Acc is accuracy, Cov is Coverage (%), Sens is Sensitivity and Spec. is Specificity.

3.6. Comparison with state-of-the-art

A comparison between the proposed approach and representative state-of-the-art methods is reported in Table 4. The selected works were chosen to ensure a fair and meaningful comparison, considering studies addressing seizure detection in comparable clinical settings and evaluated on the same dataset (CHB-MIT) using comparable performance metrics. In addition, the selected methods reflect different methodological strategies, including deep learning architectures [25,19], ensemble-based approaches [23], and techniques focusing on robustness, interpretability, or uncertainty estimation [14]. To further

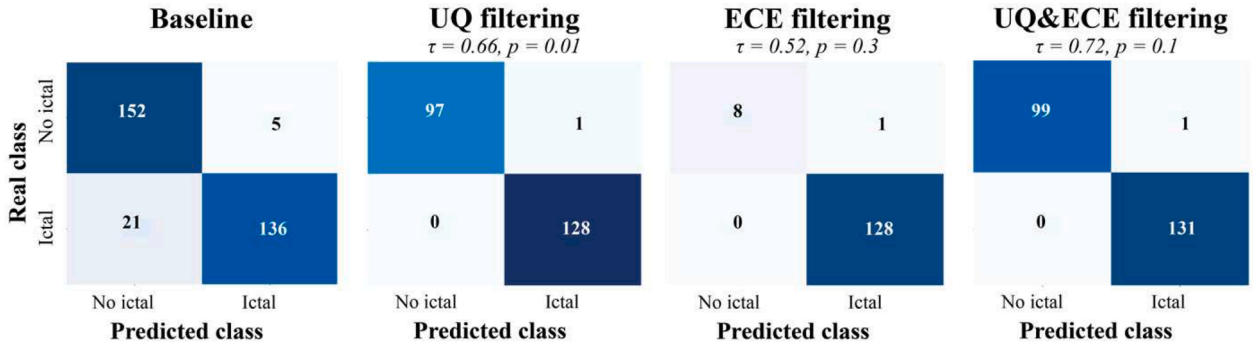


Fig. 5. Confusion matrices for the baseline model and the three filtering strategies.

Table 4
Performance comparison between our method and state-of-the-art.

Model	Input Data	Dataset	UQ	Acc.	Sens.	Spec.
Abdelhameed et al. [19]	Raw EEG (4 s epochs)	CHBMIT	No	0.98	0.98	0.98
Khan et al. [24]	EEG feature vectors	CHBMIT	No	0.92	0.94	0.91
Amrani et al. [25]	Raw EEG (1 s epochs)	CHBMIT	No	0.92	0.95	0.90
Wong et al. [14]	Raw EEG (1 s epochs)	CHBMIT	Yes	0.88	0.82	0.85
Sirpal et al. [23]	EEG feature vectors	CHBMIT	No	0.98	0.76	0.84
Proposed method UQ&ECE	Raw EEG (4 s epochs)	CHBMIT	Yes	0.996	1.00	0.99

Acc is accuracy, Sens is Sensitivity and Spec. is Specificity.

contextualize the comparison, Table 4 also highlights whether the compared methods explicitly incorporate uncertainty quantification strategies.

The integrated UQ&ECE filtering strategy (p = 0.1) achieved an accuracy of 0.996, with perfect sensitivity (1.00) and high specificity (0.99), outperforming most existing deep learning approaches on the CHB-MIT dataset. Compared with Abdelhameed et al. [19] and Sirpal et al. [23], which reported accuracies of 0.988 and 0.98, respectively, the proposed method shows a clear improvement in overall reliability. Similarly, the approaches proposed by Khan et al. [24] and Amrani et al. [25] report lower accuracy values (approximately 0.92) compared with the proposed framework. Notably, models such as Wong et al. [14], Peh et al. [25], and Sirpal et al. [23] exhibit substantially lower sensitivity (0.76–0.89) and specificity (0.62–0.85), highlighting the stronger balance achieved by our approach. Although the proposed framework applies selective filtering by identifying uncertain samples for further clinical evaluation (coverage = 74%), it achieves high performance on the retained subset of predictions. However, direct comparison with methods that classify all samples should be interpreted with caution, since selective classification inherently excludes the most uncertain cases.

4. Discussion

This study presented an integrated calibration–uncertainty framework to improve the reliability of deep learning models for EEG-based seizure detection. The results demonstrated that analyzing both calibration and predictive uncertainty provides complementary insights into model behavior, enabling the identification of an optimal configuration that balances confidence alignment and uncertainty separability.

The analysis of dropout stochasticity revealed that model reliability cannot be assessed from accuracy alone. While low dropout rates improved uncertainty separability, allowing a clear distinction between correct and incorrect predictions, higher dropout rates enhanced calibration but reduced discriminative power. These opposite trends highlight the intrinsic trade-off between calibration and uncertainty, consistent with patterns qualitatively observed in previous biomedical and EEG studies [15,16]. Our framework extends this evidence by introducing a quantitative criterion to balance these two dimensions of reliability in a clinically oriented setting. Using this approach, we identified $p = 0.1$ as the optimal configuration, corresponding to the most balanced reliability profile. This criterion reduces the subjectivity associated with qualitative visual inspection and provides a reproducible strategy for integrating calibration and uncertainty metrics.

From a methodological standpoint, the use of the OVL proved effective in quantifying the separability of uncertainty distributions between correctly and incorrectly classified samples. Compared with traditional qualitative analysis, this metric provides a quantitative indicator of model robustness. The ECE, on the other hand, offered a complementary view by assessing the agreement between model confidence and empirical accuracy. Together, these measures provide a more comprehensive characterization of model reliability.

In our work, the weighting factors were set to $w_{ECE} = 0.4$ and $w_{OVL} = 0.6$, giving slightly higher relevance to uncertainty separability for clinical reliability. These weights can be adjusted according to specific clinical requirements. For instance, increasing w_{ECE} when deployment contexts prioritize confidence calibration, or emphasizing w_{OVL} when reliable uncertainty detection is critical for patient safety. A sensitivity analysis of the weighting factors (Table 5) showed that the optimal dropout configuration depends on the relative importance assigned to calibration and uncertainty. In particular, the selected value $p = 0.1$ remains stable for configurations that moderately prioritize uncertainty, while different weightings, such as more balanced or calibration-oriented settings, lead to alternative optimal values. This behavior is consistent with the complementary roles of ECE and OVL, and highlights that the proposed framework is inherently task-dependent. Rather than relying on a single universal configuration, the method can be adapted to different clinical requirements by tuning the relative contribution of calibration and uncertainty.

The selective classification analysis confirmed that filtering based on uncertainty information can substantially improve the reliability of EEG classifiers. All filtering strategies successfully removed false negatives, the most critical type of error in clinical practice, as they correspond to missed seizure events. However, the calibration-based filtering (ECE-only) also eliminated a larger proportion of true negatives, leading to

Table 5
sensitivity analysis of optimal dropout configuration to weighting factors.

	w_{OVL}	Selected p
0.20	0.80	0.10
0.40	0.60	0.10
0.50	0.50	0.05
0.60	0.40	0.2
0.80	0.20	0.3

reduced coverage. In contrast, the uncertainty-driven and integrated approaches preserved a higher number of correct predictions while discarding unreliable samples, achieving a more favorable trade-off between accuracy and coverage. The integrated configuration, which combines both calibration and uncertainty information, yielded the best overall performance: it maintained high accuracy (0.996) and balanced class-wise F1-scores (0.99 for both classes) while retaining 74% of the data. These findings confirm that combining the two dimensions of reliability, calibration, and uncertainty leads to models that are not only more accurate but also more trustworthy and stable across varying data conditions. This behavior suggests that uncertainty-based filtering primarily identifies ambiguous segments, which may be more appropriately handled through additional clinical inspection rather than by forcing a potentially unreliable automatic classification.

While the performance gains of the integrated approach over uncertainty-only filtering may appear modest in terms of absolute metrics, several key advantages justify its adoption. In particular, the integrated configuration achieves the same accuracy as the UQ-only approach while retaining a larger proportion of samples (74%vs 72%), thereby reducing the number of segments requiring manual review. At the same time, the integration of calibration does not introduce additional computational overhead, as it relies on the same Monte Carlo Dropout inference used for uncertainty estimation. Beyond these practical considerations, the proposed framework jointly optimizes complementary aspects of model reliability by combining global confidence calibration with local uncertainty discrimination. This dual perspective provides a more comprehensive characterization of model behavior and offers a principled, reproducible strategy for selecting the optimal dropout configuration. Even when performance gains are moderate, these properties support the overall value of the integrated approach.

Despite these advances, some limitations remain. First, the proposed framework should be tested on alternative architectures, such as simpler CNN-based models, to verify that the benefits of the integrated calibration–uncertainty strategy are architecture-agnostic. Second, the current implementation relies on a fixed dropout configuration, including both dropout rate and placement within the network. Although this design follows standard model selection practices, exploring alternative dropout placements and architectural stochasticity could further improve robustness, especially in limited-data scenarios. Furthermore, future research could explore alternative uncertainty quantification techniques, including Deep Ensembles and Bayesian models [26], to compare their ability to capture the predictive uncertainty and to assess their computational feasibility for clinical deployment. An additional limitation concerns the relatively low number of MCD forward passes ($N = 10$) used for uncertainty estimation [20]. This choice was made to reduce computational cost, although higher N values could further stabilize uncertainty estimates and will be explored in future work. In addition, the interictal segment selection strategy adopted in this study was intentionally conservative, as non-seizure segments were required to be at least 30 s away from seizure events. While this choice reduced ambiguity in peri-ictal regions and helped limit label noise, it excluded close-to-boundary segments and therefore does not fully reflect real-time clinical deployment, where transitional EEG activity must also be classified. Another limitation concerns the segment-level formulation of the proposed framework, which does not preserve the temporal order of EEG segments and therefore does not explicitly enforce temporal consistency across consecutive predictions. As a result, the continuity of seizure events cannot be guaranteed. Addressing this limitation would require extending the analysis to the event level, enabling the assessment of temporal coherence and the reliable detection of complete seizure episodes. Finally, a more extensive cross-dataset validation and subject-specific evaluation should be performed to assess generalization across patient cohorts, acquisition protocols, and recording conditions. In particular, validating the framework on external datasets characterized by different demographic and clinical distributions (e.g., pediatric versus adult populations) would provide a more comprehensive assessment of its ability to handle cross-domain variability in real-world clinical scenarios.

5. Conclusion

This study introduced an integrated calibration–uncertainty framework to improve the reliability of deep learning models for EEG-based seizure detection. By jointly analyzing model calibration and predictive uncertainty, the proposed pipeline enables the identification of an optimal operating configuration and the implementation of selective classification based on uncertainty thresholds. The integrated approach achieved an accuracy of 99.6% (+8% compared with the baseline) with a coverage of 74%, outperforming models optimized for calibration or uncertainty alone and demonstrating a more stable and reliable classification behavior. Overall, the framework provides a quantitative and reproducible strategy for building robust and calibrated AI systems in EEG analysis, representing a practical step toward clinically reliable deep learning applications in neurophysiology.

Ethics statement

The study was conducted using publicly available, fully anonymized datasets. No new data were collected, and no experiments involving human participants were performed by the authors.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to improve the readability and language of manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

CRediT authorship contribution statement

Silvia Seoni: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Filippo Molinari:** Writing – review & editing, Validation, Methodology, Conceptualization. **Margherita Benevieri:** Writing – review & editing, Formal analysis, Data curation. **Massimo Salvi:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T.A. Milligan, Epilepsy: a clinical overview, Am. J. Med. 134 (7) (2021) 840–847, https://doi.org/10.1016/J.AMJMED.2021.01.038.
[2] U.R. Acharya, S. Vinitha Sree, G. Swapna, R.J. Martis, J.S. Suri, Automated EEG analysis of epilepsy: a review, Knowl. Based. Syst. 45 (2013) 147–165, https://doi.org/10.1016/J.KNOSYS.2013.02.014.
[3] L.M. Patnaik, O.K. Manyam, Epileptic EEG detection using neural networks and post-classification, Comput. Methods Programs Biomed. 91 (2) (2008) 100–109, https://doi.org/10.1016/J.CMPB.2008.02.005.
[4] S. Wong, et al., Channel-annotated deep learning for enhanced interpretability in EEG-based seizure detection, Biomed. Signal Process. Control 103 (2025) 107484, https://doi.org/10.1016/J.BSPC.2024.107484.
[5] T. Shawly, A.A. Alsheikhy, Eeg-based detection of epileptic seizures in patients with disabilities using a novel attention-driven deep learning framework with SHAP interpretability, Egypt. Inform. J. 31 (2025) 100734, https://doi.org/10.1016/J.EIJ.2025.100734.
[6] G. Amrani, A. Adadi, M. Berrada, An explainable hybrid DNN model for seizure vs. Non-seizure classification and seizure localization using multi-dimensional EEG

signals, Biomed. Signal Process. Control 95 (2024) 106322, https://doi.org/10.1016/J.BSPC.2024.106322.
[7] Y. Ding, W. Zhao, Channel selection for seizure detection based on explainable AI with Shapley values, IEEE Sens. J. 24 (16) (2024) 26126–26135, https://doi.org/10.1109/JSEN.2024.3422388.
[8] P. Sirpal, W.A. Sikora, H.H. Refai, Brain state network dynamics in pediatric epilepsy: chaotic attractor transition ensemble network, Comput. Biol. Med. 188 (2025) 109832, https://doi.org/10.1016/J.COMPBIOMED.2025.109832.
[9] H. Asgharnezhad, et al., Objective evaluation of deep uncertainty predictions for COVID-19 detection, Sci. Rep. 12 (1) (2022), https://doi.org/10.1038/s41598-022-05052-x.
[10] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks [Online]. Available: http://arxiv.org/abs/1706.04599, 2017.
[11] S. Seoni, V. Jahmunah, M. Salvi, P.D. Barua, F. Molinari, U.R. Acharya, Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013–2023), Comput. Biol. Med. 165 (2023) 107441, https://doi.org/10.1016/J.COMPBIOMED.2023.107441.
[12] Ó.W. Gómez-Morales, S. Escalante-Escobar, D.F. Collazos-Huertas, A.M. Álvarez-Meza, G. Castellanos-Dominguez, Uncertainty-aware deep learning for robust and interpretable MI EEG using channel dropout and LayerCAM integration, Appl. Sci. (Switzerland) 15 (14) (2025) 8036, https://doi.org/10.3390/APP15148036.
[13] Y. Geifman, R. El-Yaniv, Selective classification for Deep Neural networks, Adv. Neural Inf. Process. Syst. 2017-December (2017) 4879–4888. Accessed: Aug. 01, 2025. [Online]. Available: https://arxiv.org/pdf/1705.08500.
[14] S. Wong, A. Simmons, J.R. Villicana, S. Barnett, Estimating patient-level uncertainty in seizure detection using group-specific out-of-distribution detection technique, Sensors 23 (20) (2023) 8375, https://doi.org/10.3390/S23208375.
[15] H.U. Jiahao, M.M. Ur Rahman, T. Al-Naffouri, T.M. Laleg-Kirati, Uncertainty estimation and model calibration in EEG signal classification for epileptic seizures detection, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2024, 2024, https://doi.org/10.1109/EMBC53108.2024.10782858.
[16] P.S. Nzakuna, V. Gallo, V. Paciello, A. Lay-Ekuakille, A.K. Lusala, Monte Carlo-based strategy for assessing the impact of EEG data uncertainty on confidence in convolutional neural network classification, IEEE Access 13 (2025) 85342–85362, https://doi.org/10.1109/ACCESS.2025.3570134.
[17] Guttag, J. (2010). CHB-MIT Scalp EEG Database (version 1.0.0). *PhysioNet*. RRID: SCR_007345. https://doi.org/10.13026/C2K01R Available: https://physionet.org/content/chbmit/1.0.0/.
[18] A. Shoeb, H. Edwards, J. Connolly, B. Bourgeois, S.T. Treves, J. Guttag, Patient-specific seizure onset detection, Epilep. Behav. 5 (4) (2004) 483–498, https://doi.org/10.1016/J.YEBEH.2004.05.005.
[19] A. Abdelhameed, M. Bayoumi, A deep learning approach for automatic seizure detection in children with epilepsy, Front. Comput. Neurosci. 15 (2021) 650050, https://doi.org/10.3389/FNCOM.2021.650050/BIBTEX.
[20] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: representing model uncertainty in Deep learning, in: 33rd International Conference on Machine Learning, ICML 2016 3, 2015, pp. 1651–1660. Accessed: Jul. 27, 2023. [Online]. Available: https://arxiv.org/abs/1506.02142v6.
[21] A. Jungo, F. Balsiger, M. Reyes, Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation, Front. Neurosci. 14 (2020) 282, https://doi.org/10.3389/fnins.2020.00282.
[22] A. Mehrtash, W.M. Wells, C.M. Tempany, P. Abolmaesumi, T. Kapur, Confidence calibration and predictive uncertainty estimation for deep medical image segmentation, IEEE Trans. Med. Imaging 39 (12) (2020) 3868–3878, https://doi.org/10.1109/TMI.2020.3006437.
[23] P. Sirpal, W.A. Sikora, H.H. Refai, Brain state network dynamics in pediatric epilepsy: chaotic attractor transition ensemble network, Comput. Biol. Med. 188 (2025) 109832, https://doi.org/10.1016/J.COMPBIOMED.2025.109832.
[24] F.A. Khan, Z. Umar, A. Jolfaei, M. Tariq, Explainable fuzzy deep learning for prediction of epileptic seizures using EEG, IEEE Transac. Fuzzy Syst. 32 (10) (2024) 5428–5437, https://doi.org/10.1109/TFUZZ.2024.3434709.
[25] G. Amrani, A. Adadi, M. Berrada, An explainable hybrid DNN model for seizure vs. Non-seizure classification and seizure localization using multi-dimensional EEG signals, Biomed. Signal Process. Control 95 (2024) 106322, https://doi.org/10.1016/J.BSPC.2024.106322.
[26] Y. Gal, Z. Ghahramani, Bayesian convolutional Neural networks with Bernoulli approximate variational inference, Accessed: Apr. 01, 2025. [Online]. Available: https://arxiv.org/abs/1506.02158v6, 2015.