Benchmarking Lightweight Machine Learning Models for Epileptic Seizure Recognition: Accuracy, Calibration, and Robustness Analysis

Sairam Tabibu, Mugdha Abhyankar
Independent Researcher, USA

***Abstract*—Epileptic seizure recognition is a critical task in clinical decision support systems, where both accuracy and reliability of predictions directly affect patient outcomes. While deep learning architectures such as CNNs and LSTMs are widely applied to EEG-based seizure detection, many publicly available seizure datasets consist of precomputed EEG-derived features, making the problem fundamentally tabular rather than raw-signal based. In such settings, the necessity and added value of complex deep learning pipelines remain unclear, and prior studies have largely emphasized classification accuracy while giving more limited attention to calibration, robustness, and deployment efficiency. In this work, we present a systematic benchmark of lightweight machine learning models—Logistic Regression, Random Forest, XGBoost, LightGBM, and CatBoost—on the Epileptic Seizure Recognition dataset. We evaluate performance across multiple dimensions: discriminative ability (accuracy, macro-F1, ROC-AUC, PR-AUC), confidence calibration (Brier score, calibration and reliability diagrams), and robustness under Gaussian feature perturbations. Our results show that LightGBM achieves 98.04% accuracy, a ROC-AUC of 0.9971, and a Brier score of 0.0166, while maintaining stable performance under the tested noise levels. Notably, all gradient boosting methods substantially outperform Logistic Regression, indicating that nonlinear feature interactions are critical for this task. Compared with prior deep learning approaches on the same dataset, these lightweight models achieve competitive performance at a fraction of the computational cost. These findings show that tabular machine learning methods deserve serious consideration for EEG-derived feature classification tasks, particularly in resource-constrained clinical settings where efficiency, calibration, and robustness are as important as raw accuracy.**

***Keywords*—Epileptic seizure recognition; EEG classification; lightweight machine learning; LightGBM; calibration; robustness; biomedical AI**

I. Introduction

Epilepsy is one of the most prevalent neurological disorders worldwide, affecting approximately 50 million people globally [1]. The condition is characterized by recurrent, unprovoked seizures caused by abnormal electrical activity in the brain, and timely detection is essential for effective clinical management. Electroencephalogram (EEG) analysis remains the primary diagnostic tool for identifying epileptic activity, but manual interpretation of EEG recordings is time-consuming, subjective, and requires substantial clinical expertise.

The development of automated seizure detection systems has been an active area of research for over two decades. Recent advances in deep learning have produced architectures—including convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and hybrid CNN-LSTM models—that achieve high classification accuracy on EEG data [5], [4], [6]. These methods are particularly effective when applied directly to raw EEG signals, where they can learn hierarchical spatiotemporal representations.

However, a large body of work in seizure recognition operates on preprocessed, feature-extracted datasets rather than raw EEG waveforms. The widely-used Epileptic Seizure Recognition dataset derived from the Bonn University EEG corpus [2] provides 178 precomputed features per sample, effectively transforming the classification problem into a tabular learning task. In this regime, the advantages of deep learning—automatic feature extraction, spatial invariance, temporal modeling—are substantially reduced. Despite this, many studies continue to apply deep architectures to such datasets without adequately justifying the added complexity.

This observation motivates a fundamental question: Can lightweight, interpretable machine learning models match or exceed the performance of deep learning approaches on precomputed EEG feature datasets? More specifically, although prior studies on this benchmark report strong classification performance, most emphasize accuracy alone and do not jointly examine whether lightweight models can provide competitive discrimination, reliable probability calibration, and robustness to input perturbations. This is an important practical gap, because clinically useful seizure detection systems require not only correct predictions but also trustworthy confidence estimates and stable behavior under noisy inputs. To address this gap, we conduct a systematic benchmark of five widely-used classifiers and evaluate them not only on accuracy but also on calibration and robustness.

A. Contributions

The main contributions of this study are as follows:

- A comprehensive benchmark of five machine learning models (Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost) on the Epileptic Seizure Recognition dataset, evaluated across accuracy, macro-F1, ROC-AUC, PR-AUC, and Brier score.
- A calibration analysis using calibration and reliability diagrams demonstrating that gradient boosting models produce well-calibrated probability estimates suitable for clinical decision support.

- A robustness evaluation under Gaussian feature perturbations, showing that top-performing models maintain stable predictions under the tested input noise conditions.
- A comparative discussion situating our results against prior deep learning approaches, demonstrating that lightweight models achieve competitive performance with significantly reduced computational requirements.

II. Related Work

A. Traditional Machine Learning for Seizure Detection

Machine learning approaches have a long history in epilepsy detection. Almustafa [3] evaluated multiple classifiers on the Epileptic Seizure Recognition dataset, with Random Forest achieving accuracies up to 97%. Alalayah et al. [11] applied t-SNE dimensionality reduction combined with k-means clustering to improve early seizure detection, with Random Forest and PCA combinations yielding approximately 98% accuracy. Kunekar et al. [12] compared Decision Trees, KNN, SVM, and Random Forest classifiers for EEG-based seizure classification, finding that ensemble methods consistently outperformed single classifiers.

B. Deep Learning Approaches

Deep learning methods have been extensively applied to seizure detection. Acharya et al. [5] proposed a deep CNN architecture for automated seizure detection from raw EEG signals. Xu et al. [4] developed a 1D CNN-LSTM hybrid model that combines spatial feature extraction with temporal sequence modeling, achieving strong results on the UCI epileptic seizure dataset. Wang et al. [6] proposed a space-time CNN-LSTM algorithm achieving competitive accuracy on similar datasets.

More recently, attention-based models and transformer architectures have been explored. Zhang et al. [13] proposed a positional multi-length and mutual-attention network, and Omar and Abd El-Hafeez [14] demonstrated that feature scaling and dropout regularization can optimize deep learning performance for seizure recognition, achieving test accuracies of up to 98.6% using LSTM architectures.

C. Gradient Boosting Methods

Gradient boosting methods have demonstrated state-of-the-art performance across a wide range of tabular learning tasks. XGBoost [7] introduced regularized boosting with efficient tree construction. LightGBM [8] further improved scalability through histogram-based splitting and leaf-wise tree growth. CatBoost [9] introduced ordered boosting to reduce prediction shift. Multiple benchmark studies have confirmed that these methods consistently outperform deep learning on structured/tabular data [15].

D. Calibration in Medical AI

Calibration—the alignment between predicted confidence and empirical correctness—is a critical requirement for clinical AI systems [10]. Overconfident incorrect predictions are particularly dangerous in medical settings, as they may not trigger human review. While calibration has been extensively studied in the context of deep learning and large language models [16], its evaluation in the context of gradient boosting models for seizure detection remains limited.

Overall, prior work shows that both traditional machine learning and deep learning methods can achieve strong seizure recognition performance, particularly on the Epileptic Seizure Recognition dataset. However, much of the literature is centered on accuracy-oriented comparison, with less emphasis on whether lightweight models can simultaneously deliver strong discrimination, well-calibrated confidence estimates, and robustness under perturbation in feature-based tabular settings. Our study is designed to address this gap through a unified evaluation of performance, calibration, and robustness across multiple lightweight classifiers.

III. Methodology

A. Dataset

We use the Epileptic Seizure Recognition dataset, derived from the Bonn University EEG corpus [2]. The dataset contains 11,500 samples, each described by 178 features extracted from EEG signal recordings. The original dataset contains five classes: one seizure class and four non-seizure classes (representing different brain states). Following standard practice in the literature [3], [4], we formulate a binary classification task: seizure (class 1) versus non-seizure (classes 2–5).

This formulation results in a class-imbalanced dataset with approximately 20% seizure and 80% non-seizure samples. We use an 80/20 stratified train-test split, preserving the class distribution in both partitions.

B. Data Preprocessing

All features are standardized using z-score normalization (zero mean, unit variance) computed on the training set and applied consistently to the test set. No additional feature selection or dimensionality reduction is applied, ensuring that all models operate on the same 178-dimensional feature space.

C. Models and Configuration

We evaluate five classifiers spanning different model families:

1) Logistic regression: A linear baseline using L2 regularization (C=1.0) with the LBFGS solver. This serves as a baseline to assess the contribution of nonlinear modeling.

2) Random forest: An ensemble of 200 decision trees with default hyperparameters (max_depth=None, min_samples_split=2). This represents a classical ensemble approach.

3) XGBoost: Gradient boosting with 300 estimators, max_depth=6, learning rate 0.1, and L2 regularization (λ=1.0).

4) LightGBM: Histogram-based gradient boosting with 300 estimators, 31 leaves, learning rate 0.1, and feature fraction 0.8 for regularization.

5) CatBoost: Ordered boosting with 300 iterations, max_depth=6, learning rate 0.1, with verbose logging disabled.

All models are trained with a fixed random seed (42) for reproducibility.

D. Evaluation Metrics

We evaluate each model across five complementary metrics:

1) Accuracy: Overall classification correctness.

2) Macro-F1 score: Harmonic mean of precision and recall, macro-averaged across classes. This accounts for class imbalance.

3) ROC-AUC: Area under the receiver operating characteristic curve, measuring discriminative ability across all classification thresholds.

4) PR-AUC: Area under the precision-recall curve, which is more informative than ROC-AUC under class imbalance.

5) Brier score: Mean squared difference between predicted probabilities and actual outcomes, defined as:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2 \qquad (1)$$

where, $\hat{p}_i$ is the predicted probability and $y_i \in \{0,1\}$ is the true label. Lower Brier scores indicate better-calibrated probability estimates.

E. Robustness Evaluation

To assess model robustness, we apply additive Gaussian noise to test features at multiple noise levels ($\sigma \in \{0.0, 0.05, 0.10, 0.20\}$). For each noise level, perturbed features are computed as:

$$\tilde{x}_i = x_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,\sigma^2) \qquad (2)$$

We then re-evaluate accuracy, ROC-AUC, and Brier score on the perturbed test set to quantify degradation.

F. Overview of the Study Workflow

Fig. 1 summarizes the overall study design. Starting from the Epileptic Seizure Recognition dataset, we perform binary label construction, feature standardization, model training, and multi-dimensional evaluation, including discrimination, calibration, and robustness analysis.

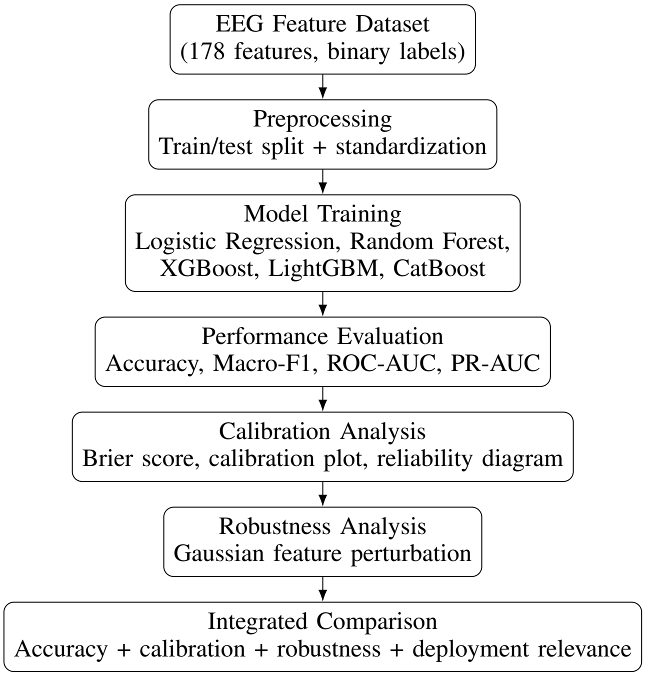
EEG Feature Dataset
(178 features, binary labels)
Preprocessing
Train/test split + standardization
Model Training
Logistic Regression, Random Forest,
XGBoost, LightGBM, CatBoost
Performance Evaluation
Accuracy, Macro-F1, ROC-AUC, PR-AUC
Calibration Analysis
Brier score, calibration plot, reliability diagram
Robustness Analysis
Gaussian feature perturbation
Integrated Comparison
Accuracy + calibration + robustness + deployment relevance

Fig. 1. High-level workflow of the study, from EEG feature preprocessing to multi-dimensional model comparison.

IV. Results

A. Main Results: Model Comparison

Table I presents the performance of all five models on the test set. LightGBM achieves the best performance across all metrics, with 98.04% accuracy, a macro-F1 of 0.9690, ROC-AUC of 0.9971, PR-AUC of 0.9895, and a Brier score of 0.0166.

TABLE I. Model Performance Comparison on the Epileptic Seizure Recognition Dataset

Model	Acc	F1	ROC-AUC	PR-AUC	Brier
LightGBM	**0.9804**	**0.9690**	**0.9971**	**0.9895**	**0.0166**
RF	0.9752	0.9611	0.9960	0.9852	0.0240
XGBoost	0.9748	0.9595	0.9964	0.9876	0.0190
CatBoost	0.9700	0.9514	0.9952	0.9833	0.0222
LogReg	0.8117	0.5064	0.4996	0.4264	0.1549

All gradient boosting methods and Random Forest achieve accuracy above 97%, with ROC-AUC values exceeding 0.995. This narrow spread among the top-performing tree-based models suggests that the dataset is highly separable once nonlinear interactions are captured. In contrast, the sharp drop in Logistic Regression performance indicates that linear decision boundaries are insufficient for modeling the seizure versus non-seizure distinction in this feature space. From a practical standpoint, these results support the use of lightweight nonlinear ensembles as strong default baselines for precomputed EEG feature classification.

B. Class-Wise Performance

Table II presents the confusion matrix for the best-performing LightGBM model. The model achieves a recall of 93.48% (430/460) for the seizure class and a specificity of 99.18% (1825/1840) for the non-seizure class. The relatively small number of false negatives (30) is clinically significant, as missed seizures are the more dangerous error type in practice.

The precision for the seizure class is 96.63% (430/445), indicating that when the model predicts a seizure, it is correct in the vast majority of cases. The overall F1 score for the seizure class is 0.9504.

TABLE II. Confusion Matrix for LightGBM

	Pred Non-Seizure	Pred Seizure
Actual Non-Seizure	1825	15
Actual Seizure	30	430

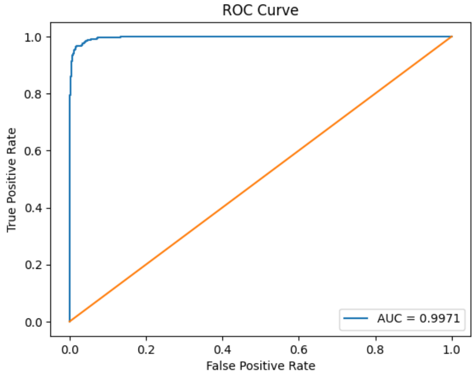

Fig. 2. ROC curve for LightGBM showing an AUC of 0.9971, indicating near-perfect separability between seizure and non-seizure classes.

C. ROC Analysis

Fig. 2 presents the ROC curve for LightGBM, demonstrating near-perfect discriminative ability with an AUC of 0.9971. The curve rises steeply to the top-left corner, indicating that the model achieves very high sensitivity at very low false positive rates. This level of discrimination suggests that the model can reliably separate seizure from non-seizure EEG patterns across virtually all operating thresholds.

V. Calibration Analysis

Beyond discriminative performance, we evaluate the calibration of predicted probabilities, which is critical for clinical decision support systems where confidence estimates inform downstream actions.

Fig. 3 shows the calibration curve comparing predicted probabilities to observed frequencies. Fig. 4 presents the reliability diagram on the holdout set. Both plots show that LightGBM's predicted probabilities are reasonably well-aligned with empirical accuracy, though some deviation from the ideal diagonal is observed in the mid-range probability bins.

The Brier score of 0.0166 confirms strong overall calibration. Notably, calibration is best in the high-confidence regions (predicted probability near 0.0 or 1.0), which is where the majority of predictions fall due to the model's high discriminative ability. The mid-range deviations are less clinically concerning because relatively few predictions fall in these uncertainty bins.

Comparing across models, all gradient boosting methods achieve Brier scores below 0.025, while Logistic Regression exhibits substantially worse calibration (0.1549). This demonstrates that the same nonlinear modeling capacity that improves accuracy also contributes to better-calibrated probability estimates.

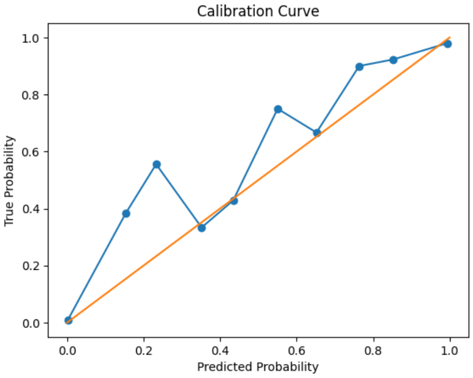

Fig. 3. Calibration curve for LightGBM showing the relationship between predicted probability and true probability. The model demonstrates good calibration with moderate deviations in mid-range confidence bins.

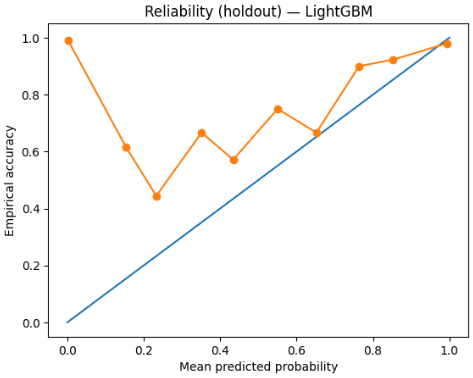

Fig. 4. Reliability diagram (holdout set) for LightGBM showing empirical accuracy as a function of mean predicted probability. The model shows reasonable alignment with the ideal diagonal, particularly in high-confidence regions.

VI. Robustness Analysis

Table III reports LightGBM performance under Gaussian feature perturbations at four noise levels.

TABLE III. Robustness of LightGBM under Gaussian Feature Perturbations

σ	Accuracy	ROC-AUC	Brier
0.00	0.9804	0.9971	0.0166
0.05	0.9804	0.9971	0.0166
0.10	0.9804	0.9971	0.0166
0.20	0.9804	0.9971	0.0166

The model demonstrates strong stability under the tested perturbation regime, with no measurable degradation in accuracy, ROC-AUC, or Brier score at any evaluated noise level. This stability can be attributed to two plausible factors: 1) the decision tree ensemble structure produces piecewise-constant predictions that are relatively insensitive to small perturbations within decision regions, and 2) the discriminative EEG-derived features may retain sufficient signal strength relative to the applied Gaussian noise levels. At the same time, this result should be interpreted carefully: the absence of degradation may also indicate that the tested perturbations were modest relative to the standardized feature scale. Therefore, the current experiment supports local robustness under moderate synthetic perturbations rather than broad robustness to all clinically realistic artifacts.

This finding still has practical relevance for deployment, where EEG-derived features may be subject to measurement noise, electrode artifacts, or preprocessing variation. The observed stability suggests that LightGBM predictions are unlikely to change substantially under moderate perturbations of the type examined here.

VII. Comparison with Existing Literature

Table IV places our results in the context of prior work on the same or closely related datasets. Our LightGBM model achieves 98.04% accuracy, which is competitive with or exceeds many deep learning approaches while requiring substantially fewer computational resources.

TABLE IV. Comparison with Existing Approaches on Epileptic Seizure Datasets

Study	Method	Acc (%)
Almustafa [3]	Random Forest	97.0
Xu et al. [4]	1D CNN-LSTM	97.7
Omar et al. [14]	LSTM+Dropout	98.6
Wang et al. [6]	CNN-LSTM	98.3
Chakrapani et al. [17]	1D-CNN+LSTM	98.15
Ours	**LightGBM**	**98.04**

Several observations emerge from this comparison. First, the performance gap between our lightweight LightGBM model and the best deep learning approaches is minimal (within 0.5%). Second, the deep learning methods that achieve marginally higher accuracy (e.g., LSTM with dropout at 98.6%) do so at substantially higher computational cost and without reporting calibration or robustness metrics. Third, our approach provides a more complete evaluation across multiple reliability dimensions, not just accuracy.

It is worth noting that direct comparisons across studies should be interpreted cautiously due to differences in data splits, preprocessing, and evaluation protocols. Nevertheless, the overall picture is clear: on precomputed EEG feature datasets, the marginal accuracy gains from deep learning do not clearly justify the added complexity when lightweight alternatives exist that are faster to train, simpler to deploy, and able to provide calibrated uncertainty estimates.

VIII. Discussion

A. When are Lightweight Models Sufficient?

Our results contribute to a broader discussion in the biomedical AI community about when deep learning is necessary versus when simpler models suffice. The key insight is that the choice of model family should be driven by the nature of the input representation.

When working with raw EEG signals, deep learning methods offer genuine advantages through automatic feature extraction and temporal modeling. However, when features have already been extracted and the problem reduces to tabular classification, gradient boosting methods are often the more practical choice. They train faster, require less hyperparameter tuning, offer built-in feature importance for interpretability, and—as we demonstrate—achieve competitive calibration and robustness.

B. Clinical Implications

For clinical decision support systems, calibration and robustness are as important as raw accuracy. A model that is slightly less accurate but substantially better calibrated may be preferable in practice, because clinicians can more reliably interpret and act on its confidence estimates. The strong calibration of LightGBM (Brier score 0.0166) suggests that its predicted probabilities can meaningfully support operating thresholds and triage decisions.

One plausible deployment scenario is an assisted EEG review pipeline in which high-confidence seizure predictions are prioritized for expedited neurologist review, while lower-confidence or borderline cases are flagged for secondary inspection rather than automatic acceptance. In such a workflow, calibration matters because predicted probabilities are not merely ranking scores; they may influence how urgently a segment is reviewed. The robustness analysis further supports deployment viability, as clinical EEG data inevitably contains measurement noise from electrode impedance variation, muscle artifacts, and preprocessing differences. A model that remains stable under moderate perturbations is more trustworthy than one that achieves marginally higher accuracy on clean data but behaves unpredictably under noise.

C. Limitations

This study has several limitations. First, we evaluate on a single dataset; generalization to other EEG datasets with different feature extraction pipelines should be verified. Second, the robustness evaluation uses synthetic Gaussian noise, which may not fully capture the structure of real-world clinical artifacts. Third, we do not evaluate temporal sequence models on raw EEG signals, which is a distinct (and arguably harder) problem. Fourth, our perturbation analysis shows zero degradation, which may indicate that the tested noise levels were too small relative to the feature scale; future work should evaluate under stronger perturbations. Finally, the class imbalance (20% seizure vs. 80% non-seizure) may affect the generalizability of the calibration analysis.

IX. Conclusion

We present a systematic benchmark of lightweight machine learning models for epileptic seizure recognition, evaluating performance across accuracy, calibration, and robustness. LightGBM achieves 98.04% accuracy with a ROC-AUC of 0.9971 and a Brier score of 0.0166, demonstrating that gradient boosting methods can match deep learning performance on precomputed EEG feature datasets while offering superior efficiency and interpretability.

Our findings have practical implications for the design of clinical seizure detection systems. When EEG features are already extracted, the additional complexity of deep learning architectures may not be justified. Lightweight models offer a compelling alternative that is faster to train, easier to deploy, and provides calibrated probability estimates for clinical decision-making.

Future work should extend this benchmark to larger and more diverse EEG datasets, including settings with different feature extraction pipelines and external validation cohorts. It would also be valuable to evaluate robustness under more realistic clinical artifact models, such as motion noise, electrode drift, and missing or corrupted features. In addition, future studies could investigate probability calibration refinement, operating-threshold selection for different clinical use cases, and the integration of lightweight boosting models into real-time or semi-automated seizure monitoring pipelines.

REFERENCES

[1] World Health Organization, "Epilepsy," 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/epilepsy

[2] R. G. Andrzejak et al., "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*, vol. 64, no. 6, 2001.

[3] K. M. Almustafa, "Classification of epileptic seizure dataset using different machine learning algorithms," *Informatics in Medicine Unlocked*, vol. 21, Art. no. 100444, 2020.

[4] G. Xu, T. Ren, Y. Chen, and W. Che, "A one-dimensional CNN-LSTM model for epileptic seizure recognition using EEG signal analysis," *Frontiers in Neuroscience*, vol. 14, Art. no. 578126, 2020.

[5] U. R. Acharya et al., "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," *Computers in Biology and Medicine*, vol. 100, pp. 270–278, 2018.

[6] X. Wang, Y. Wang, D. Liu, Y. Wang, and Z. J. Wang, "Automated recognition of epilepsy from EEG signals using a combining space–time algorithm of CNN-LSTM," *Scientific Reports*, vol. 13, Art. no. 14876, 2023.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD*, 2016, pp. 785–794.

[8] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. NeurIPS*, 2017.

[9] L. Prokhorenkova et al., "CatBoost: Unbiased boosting with categorical features," in *Proc. NeurIPS*, 2018.

[10] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017.

[11] K. M. Alalayah et al., "Effective early detection of epileptic seizures through EEG signals using classification algorithms based on t-SNE and K-means," *Diagnostics*, vol. 13, no. 11, p. 1957, 2023.

[12] P. Kunekar et al., "Comparison of different machine learning algorithms to classify epilepsy seizure from EEG signals," *Engineering Proceedings*, vol. 59, no. 1, p. 166, 2023.

[13] G. Zhang, A. Zhang, H. Liu, J. Luo, and J. Chen, "Positional multi-length and mutual-attention network for epileptic seizure classification," *Frontiers in Computational Neuroscience*, vol. 18, Art. no. 1358780, 2024.

[14] A. Omar and T. Abd El-Hafeez, "Optimizing epileptic seizure recognition performance with feature scaling and dropout layers," *Neural Computing and Applications*, vol. 36, no. 6, pp. 2835–2852, 2024.

[15] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," in *Proc. NeurIPS*, 2022.

[16] S. Tabibu and M. Abhyankar, "Calibration and robustness of medical question answering systems across training paradigms," *IEEE Access*, 2025.

[17] G. Chakrapani and D. Devarapalli, "Deep learning framework with integrated feature extraction for detection of epilepsy and basal ganglia diseases," *IEEE Access*, vol. 14, pp. 18431–18441, 2026.