Modular Machine Unlearning for Seizure Detection Using SISA Framework and Influence Functions

Aravinth A[1], Jai Karaneesh S[1], Kanagaraj M[1], Aravinth B[1]
[1]Department of Information Technology
Madras Institute of Technology Campus, Anna University
Chennai, India

Dr. P. Kola Sujatha[2]
[2]Associate Professor & Supervisor
Department of Information Technology
Madras Institute of Technology Campus, Anna University
Chennai, India
kolasujatha@mitindia.edu

***Abstract*—Deep learning–based EEG seizure detection systems achieve high accuracy but raise critical privacy concerns due to their tendency to retain patient-specific neural patterns, making compliance with regulations such as the "Right to Be Forgotten" difficult. Conventional approaches require complete model retraining to remove patient data, which is computationally expensive and impractical for real-world clinical deployment. This paper proposes a modular machine unlearning framework that enables efficient and selective removal of patient data while preserving seizure detection performance. The framework integrates two complementary unlearning strategies: the Sharded, Isolated, Sliced, and Aggregated (SISA) framework and Influence Function–based unlearning. SISA partitions the UPenn–Mayo EEG dataset into independent shards, allowing localized retraining when data removal is requested, while Influence Functions provide a retraining-free alternative by estimating and reversing the contribution of specific samples to model parameters through Hessian–vector products. To further enhance unlearning reliability, a knowledge-aware sharding encoder is introduced to distribute similar patient data across different shards, preventing residual information leakage. The framework is evaluated using EEGNet, Temporal Convolutional Networks (TCN), CNN-LSTM, and CNN-RNN architectures, with EEGNet and TCN demonstrating the most robust performance. Experimental results show that the proposed approach effectively supports both modular retraining and retraining-free unlearning while maintaining stable AUC, enabling privacy-compliant seizure detection systems for healthcare applications.**

***Index Terms*—Machine Unlearning, Seizure detection, EEG Analysis, SISA Framework, Influence Functions, Privacy-Preserving AI**

I. Introduction

A. EEG-Based Seizure Detection and Its Importance

Epileptic seizures are sudden and abnormal bursts of electrical activity in the brain that may lead to loss of awareness, convulsions, or altered behavior. Accurate and reliable seizure detection is critical for clinical monitoring, early intervention, and long-term management of epilepsy, particularly for patients affected by chronic or drug-resistant conditions. Electroencephalography (EEG) is widely used for seizure detection because it provides a non-invasive and continuous recording of brain activity with high temporal resolution. However, EEG signals are inherently complex, high-dimensional, and exhibit significant variability across patients, electrode placements, and seizure types, making automated seizure detection a challenging task.

Recent advances in machine learning and deep learning have substantially improved EEG-based seizure detection by automatically learning discriminative temporal, spatial, and spectral features from multichannel EEG signals. Models such as convolutional and temporal architectures have demonstrated strong detection accuracy and generalization compared to traditional feature-engineering approaches. Accurate seizure detection systems can enable continuous patient monitoring, timely medical response, and improved quality of life, making them an essential component of modern clinical decision-support systems.

B. Privacy Risks in Medical Deep Learning Models

Despite their effectiveness, deep learning models trained on medical EEG data introduce serious privacy risks. EEG recordings capture patient-specific neural signatures that may be implicitly memorized by trained models. Even if raw EEG data are deleted from storage, residual traces of patient information may persist within the model parameters. This exposes systems to privacy attacks such as membership inference or model inversion, where an adversary attempts to infer whether a particular patient's data was used during training. These risks are especially concerning in healthcare environments, where EEG data directly reflect sensitive neurological conditions.

Regulatory frameworks such as the General Data Protection Regulation (GDPR) emphasize the protection of personal medical data and grant individuals the "Right to Be Forgotten," allowing them to request the removal of their data from digital systems. In the context of deep learning, this requirement extends beyond data storage and necessitates the removal of learned representations derived from patient data. Without effective mechanisms to address this issue, deployed seizure detection models may violate legal, ethical, and clinical privacy standards.

C. Limitations of Conventional Retraining-Based Forgetting

Traditional approaches to data removal rely on excluding the affected samples and retraining the entire model from scratch. While this approach ensures exact forgetting, it is computationally expensive, time-consuming, and impractical for large-scale EEG datasets or continuously deployed clinical systems. Frequent retraining disrupts model availability,

increases infrastructure costs, and reduces system stability, making it unsuitable for real-world healthcare environments where data removal requests may occur after deployment.

Furthermore, EEG datasets often exhibit strong inter-patient variability. Naively retraining models without careful data organization may lead to degraded generalization or unintended loss of clinically relevant features. These limitations highlight the need for efficient, targeted unlearning strategies that avoid full retraining while preserving detection performance.

D. Motivation for Machine Unlearning in Healthcare

Machine unlearning aims to selectively remove the influence of specific training samples from a trained model without requiring complete retraining. In healthcare applications, this capability is essential to ensure regulatory compliance, protect patient autonomy, and maintain trust in AI-assisted clinical systems. Effective unlearning mechanisms allow seizure detection models to adapt dynamically to data removal requests while remaining accurate and clinically useful.

To address this need, this work investigates two complementary machine unlearning strategies. The SISA (Sharded, Isolated, Sliced, and Aggregated) framework enables modular retraining by partitioning data into independent shards, allowing only the affected shard to be retrained during unlearning. In contrast, Influence Function–based unlearning operates directly on trained model parameters, estimating and reversing the contribution of specific samples without retraining. Together, these approaches offer distinct trade-offs between accuracy, computational cost, and unlearning efficiency.

E. Contributions of This Paper

The main contributions of this work are summarized as follows:

- A hybrid machine unlearning framework for EEG-based seizure detection that integrates SISA-based modular retraining with Influence Function–based retraining-free unlearning.
- A knowledge-aware sharding encoder that leverages patient similarity analysis to distribute correlated EEG profiles across shards, preventing residual memory leakage during unlearning.
- A comprehensive experimental evaluation using multiple deep learning architectures, including EEGNet, TCN, CNN-LSTM, and PCT-Net, demonstrating effective unlearning with minimal impact on seizure detection performance.

II. Related Work

A. Deep Learning for EEG Seizure Detection

Deep learning techniques have been widely adopted for automated EEG-based seizure detection due to their ability to learn complex temporal and spatial patterns directly from raw signals. Convolutional Neural Networks (CNNs) such as EEGNet have demonstrated strong performance by extracting spatial–spectral features across EEG channels, while Temporal Convolutional Networks (TCN) and recurrent architectures such as LSTM-based models effectively capture long-range temporal dependencies in seizure evolution. These models significantly outperform traditional feature-engineering approaches; however, they inherently learn patient-specific neural signatures, which introduces privacy risks when deployed in clinical environments.

B. Machine Unlearning Approaches

Machine unlearning has emerged as a solution to address privacy concerns by enabling the removal of specific data samples from trained models. Existing unlearning methods can be broadly categorized into retraining-based and approximate unlearning techniques.

1) Retraining-Based Unlearning: Retraining-based approaches remove target data and retrain the model using the remaining dataset to ensure exact forgetting. While this guarantees strong unlearning correctness, it is computationally expensive and impractical for large-scale EEG datasets or real-time healthcare systems. Frequent retraining also disrupts continuous model availability, limiting its applicability in clinical deployments.

2) Approximate Unlearning: Approximate unlearning methods aim to reduce computational cost by modifying model parameters without full retraining. These techniques trade exactness for efficiency and are particularly suitable for scenarios requiring rapid data removal. However, approximate methods must be carefully designed to avoid instability and excessive performance degradation, especially in non-convex deep learning models.

C. SISA Framework in Privacy-Preserving Machine Learning

The Sharded, Isolated, Sliced, and Aggregated (SISA) framework introduces modular training by partitioning datasets into independent shards and training separate sub-models. When data must be forgotten, only the affected shard is retrained, significantly reducing computational overhead. Although SISA has been explored in generic machine unlearning settings, its direct application to EEG-based seizure detection remains limited, particularly in addressing patient similarity and residual information leakage across shards.

D. Influence Functions for Model Interpretability and Forgetting

Influence Functions provide a principled approach to estimate the impact of individual training samples on model parameters by approximating second-order gradients. Originally proposed for model interpretability, Influence Functions have recently been applied to approximate unlearning by reversing the estimated contribution of target samples. While promising, their application to deep EEG models requires architectural and regularization adjustments to ensure numerical stability and reliable forgetting.

E. Research Gap

Despite growing interest in privacy-preserving learning, existing works primarily focus on either modular retraining or approximate unlearning in isolation. Moreover, limited attention has been given to EEG-specific challenges, such as patient similarity and residual neural pattern leakage during unlearning. *Existing works do not jointly address efficient unlearning and EEG-specific patient similarity leakage.*

III. Problem Formulation

The seizure detection task considered in this work involves learning a mapping from multichannel EEG time-series signals to a binary output representing seizure and non-seizure states. An EEG segment can be represented as

$$X = \{x(t)\}_{t=1}^{T}, \quad x(t) \in \mathbb{R}^{C}, \tag{1}$$

where C denotes the number of EEG channels and T denotes the number of time samples. The objective is to learn a function $f_\theta(\cdot)$ parameterized by θ that maps an EEG segment to a seizure probability:

$$\hat{y} = f_\theta(X), \quad \hat{y} \in [0, 1], \tag{2}$$

where $\hat{y}$ represents the predicted probability of a seizure. While deep learning models can learn complex temporal, spatial, and spectral representations from EEG data, these learned parameters may implicitly encode patient-specific neural signatures, introducing privacy risks in medical applications.

In regulated healthcare environments, seizure detection systems must satisfy strict privacy and data protection requirements. Under regulations such as the "Right to Be Forgotten," patients may request the removal of their data from trained models. Effective unlearning requires that the influence of removed data be eliminated not only from the training dataset but also from the internal parameters of the deployed model. At the same time, the unlearning process must preserve the predictive performance of the model on the remaining data while avoiding excessive computational overhead that would hinder real-world deployment.

Formally, let a model be trained on a dataset $D = \{(X_i, y_i)\}_{i=1}^{N}$, producing parameters θ_D. Given a deletion request for a specific sample $x \in D$, the reduced dataset is defined as

$$D' = D \setminus \{x\}. \tag{3}$$

Machine unlearning aims to compute an updated parameter set $\theta_{D'}$ such that the resulting model is statistically indistinguishable from a model trained directly on D', i.e.,

$$\theta_{D'} \approx \arg\min_{\theta} \mathcal{L}(D', \theta), \tag{4}$$

without requiring full retraining from scratch on D'.

The threat model considered in this work assumes that an adversary may attempt to infer whether a particular patient's data was included during training by exploiting residual information embedded in the model parameters. Such threats include membership inference attacks, which attempt to determine whether a given sample belongs to the training set, and residual memory leakage arising from similarities between patients' EEG patterns. Therefore, an effective unlearning framework must ensure that the contribution of removed data satisfies

$$\mathcal{I}(x, \theta_{D'}) \to 0, \tag{5}$$

where $\mathcal{I}(\cdot)$ denotes the influence of a sample on the trained model, while maintaining stable seizure detection performance for the remaining population.

IV. Dataset Description and Preprocessing

This work utilizes the UPenn–Mayo intracranial EEG (iEEG) seizure detection dataset, consisting of recordings from both human patients and canine subjects. The dataset is organized into training and testing segments, where each training sample is a 1-second EEG clip labeled as either ictal (seizure) or interictal (non-seizure). Ictal segments span the full duration of seizure events, while interictal segments are randomly selected from non-seizure periods occurring at least one hour before or after any seizure onset. EEG data are stored in MATLAB `.mat` files containing multichannel signals arranged as electrodes $\times$ time, along with metadata including sampling frequency, channel names, and seizure latency for ictal samples. Human recordings were obtained using depth and subdural electrodes from patients with temporal and extratemporal lobe epilepsy, with sampling rates ranging from 500 Hz to 5000 Hz, while canine recordings were acquired using implanted subdural electrodes at 400 Hz.

Due to the large scale and heterogeneity of the raw dataset, a multi-stage preprocessing pipeline was applied. Unlabeled and unnecessary files were first removed, reducing storage overhead. All signals were then converted to a compressed `.npz` format with 32-bit floating-point precision to improve I/O efficiency. Each EEG segment was standardized to a fixed length of 400 samples and normalized on a per-channel basis using Z-score normalization:

$$\hat{x}_i(t) = \frac{x_i(t) - \mu_i}{\sigma_i + 10^{-6}}, \tag{6}$$

where μ_i and σ_i denote the mean and standard deviation of channel i, respectively.

To address variability in electrode count across subjects, truncated singular value decomposition (SVD) was applied to project all signals onto a unified 32-channel space. Given an EEG segment X, the decomposition is defined as

$$X = U \Sigma V^{T}, \tag{7}$$

and the projected signal is obtained by retaining the top $k = 32$ components:

$$X_{\text{proj}} = W_k \cdot (X - \text{avg}(X)), \tag{8}$$

where W_k denotes the matrix of principal spatial components.

Class imbalance between ictal and interictal samples was mitigated using class weighting, oversampling, and data augmentation. Oversampling was performed to approximate

$$N'_{\text{ictal}} \approx \alpha \cdot N_{\text{interictal}}, \tag{9}$$

TABLE I
SISA SHARD CONFIGURATION AND PERFORMANCE

Model	Shards	Patients/Shard	ROC-AUC
PCT_full	1	12	0.495
PCT_4_shard	4	3	0.550
PCT_3_shard	3	4	0.515
PCT_2_shard	2	6	**0.666**

where α is a subject-dependent scaling factor.

Following preprocessing, the dataset consists of fixed-length, normalized EEG segments with a unified channel count and balanced class distribution, providing a compact and standardized input suitable for seizure detection and machine unlearning experiments.

V. PROPOSED METHODOLOGY

A. SISA-Based Modular Training

To enable efficient and selective unlearning, this work adopts the Sharded, Isolated, Sliced, and Aggregated (SISA) training framework. In SISA, the training dataset is partitioned into multiple disjoint shards, where each shard contains data from a subset of patients. Each shard is trained independently, ensuring isolation of learned parameters across shards. Within each shard, data are further divided into slices, which preserve patient-wise grouping and temporal ordering during training. This organization prevents parameter interdependence across shards and enables localized retraining.

During inference, predictions from all shard-level models are aggregated to produce the final seizure detection output. When a data deletion request is issued for a particular patient, only the shard containing that patient's data is retrained from its most recent checkpoint, while all other shards remain unchanged. This modular retraining strategy significantly reduces computational overhead compared to full retraining and makes unlearning feasible in continuously deployed clinical systems.

To study the impact of shard granularity, multiple shard configurations were evaluated using the PCT-Net architecture. Increasing the number of shards improves unlearning efficiency but reduces the amount of data per shard, potentially leading to underfitting. Conversely, fewer shards provide better generalization but increase retraining cost. Experimental results demonstrate that a two-shard configuration achieves the best trade-off between detection performance and unlearning efficiency.

B. Influence Function-Based Unlearning

Influence Function (IF)–based unlearning provides a retraining-free mechanism to remove the contribution of specific training samples from a trained model. Instead of rebuilding the model from scratch, IFs approximate how the optimal model parameters would change if the forget set had never been included during training.

The first step computes the forget gradient for the samples to be removed:

$$\mathbf{v} = \nabla_\theta \mathcal{L}(D_{\text{forget}}, \theta), \tag{10}$$

where $\mathbf{v}$ captures the direction in parameter space influenced by the forget set.

To account for the curvature of the loss landscape, the Hessian–Vector Product (HVP) is computed as:

$$\text{HVP}(p) = H \cdot p, \tag{11}$$

where H is the Hessian of the loss with respect to θ. Direct computation of H is infeasible for deep networks; therefore, automatic differentiation is used to compute HVP efficiently.

Influence estimation requires the inverse Hessian–Vector Product (iHVP), which is obtained by solving:

$$H\mathbf{x} = \mathbf{v}, \tag{12}$$

using the Conjugate Gradient method, yielding:

$$\mathbf{x} \approx H^{-1}\mathbf{v}. \tag{13}$$

Finally, the unlearned model parameters are updated as:

$$\theta_{\text{unlearned}} = \theta_{\text{original}} + \mathbf{x}, \tag{14}$$

which approximates the parameter state the model would have achieved had the forget set never been included.

C. Knowledge-Aware Sharding Encoder

A key challenge in unlearning arises when different patients exhibit highly similar EEG patterns. Removing one patient's data may be insufficient if correlated representations remain in other shards. To address this, a Knowledge-Aware Sharding Encoder is introduced to guide shard assignment.

A CNN-based encoder extracts low-dimensional embedding vectors from each patient's EEG segments, capturing essential spatio-temporal characteristics. Pairwise cosine similarity between patient embeddings is computed as:

$$\text{Similarity}(A, B) = \frac{A \cdot B}{\|A\| \, \|B\|}, \tag{15}$$

where A and B denote embedding vectors.

Hierarchical clustering is applied to the similarity matrix to identify groups of patients with correlated EEG profiles. Patients within the same cluster are deliberately assigned to different shards during SISA partitioning. This prevents residual memory leakage and ensures that unlearning one patient does not suffer from representational leakage caused by correlated samples from other patients, thereby improving the reliability and verifiability of the unlearning process.

VI. RESULTS AND DISCUSSION

A. Baseline Model Performance

Baseline seizure detection performance was first established under the SISA training framework to evaluate model stability before applying unlearning mechanisms. The results are summarized in Table II. Among the evaluated architectures, the Temporal Convolutional Network (TCN) demonstrated the most consistent and robust performance, achieving a narrow and high ROC-AUC range (0.93–0.94) across shards. This indicates strong generalization and reliable seizure discrimination despite data partitioning. EEGNet also exhibited stable

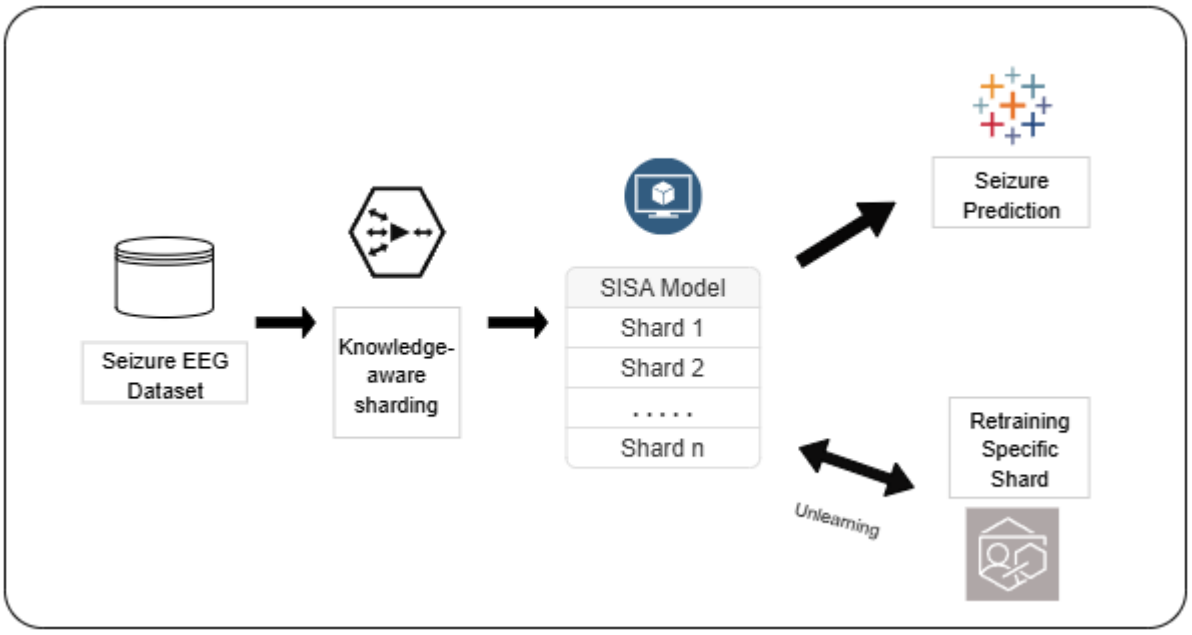

Fig. 1. SISA-based modular training and shard-level unlearning workflow.

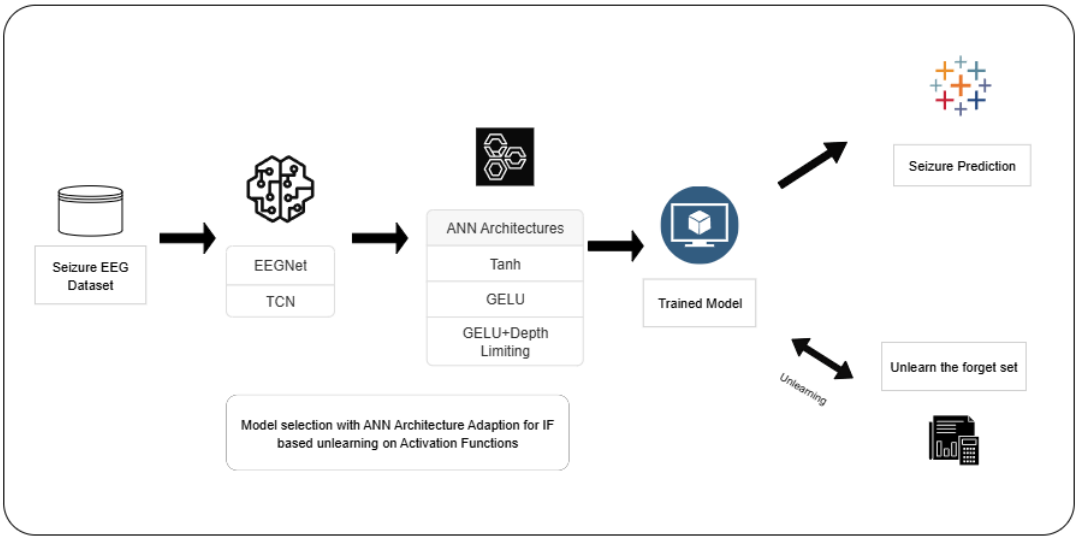

Fig. 2. Influence Function–based unlearning pipeline for EEG seizure detection.

behavior across shards, although its test AUC showed greater variability, reflecting sensitivity to shard composition.

In contrast, CNN-LSTM and CNN-RNN models achieved very high training accuracy but comparatively weaker validation performance, indicating overfitting under shard-based training. Based on these observations, EEGNet and TCN were selected as the primary architectures for subsequent unlearning experiments.

TABLE II
BASELINE MODEL PERFORMANCE UNDER SISA FRAMEWORK

Model	Shard Count	Training Accuracy	Validation Accuracy	Test AUC
EEGNet	3 Shards	0.79–0.96	0.39–0.90	0.30–0.71
TCN	3 Shards	0.72–0.83	0.52–0.89	0.93–0.94
CNN+LSTM	3 Shards	0.86–0.95	0.32–0.61	–
CNN+RNN	3 Shards	0.96–0.99	0.54–0.71	–

B. Influence Function–Based Unlearning Results

Influence Function (IF)–based unlearning was applied to the selected EEGNet and TCN models to evaluate retraining-free data removal. Table III compares model performance before and after unlearning a designated forget set. Across most configurations, ROC-AUC remained stable or improved after unlearning, indicating that the overall discriminative capability

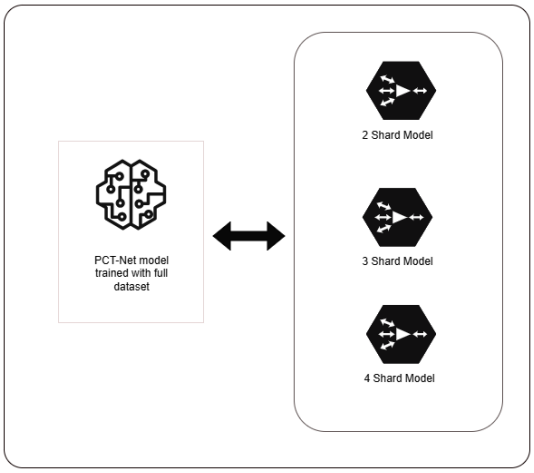

Fig. 3. Knowledge-aware sharding using CNN-based patient embeddings and similarity-driven shard assignment.

of the models was preserved.

In the EEGNet–Balanced configuration, a substantial reduction in F1-score and Recall was observed after unlearning, demonstrating successful targeted forgetting of seizure patterns associated with the removed patient data. At the same time, the stability of ROC-AUC confirms that global model utility was not compromised. For TCN models, IF-based unlearning showed scenario-dependent behavior, reflecting differences in how strongly the forgotten data influenced model parameters. Importantly, all IF-based unlearning was achieved without retraining, highlighting its computational efficiency.

C. SISA Shard Optimization Results

To analyze the trade-off between unlearning efficiency and detection performance, the PCT-Net model was trained under multiple SISA shard configurations. The results are shown in Table IV. While the single-shard baseline achieved the highest accuracy, its low ROC-AUC indicates weak generalization. Increasing the number of shards reduced accuracy due to data fragmentation.

The two-shard configuration achieved the highest ROC-AUC (0.666), demonstrating the best balance between generalization and modular retraining efficiency. This configuration enables unlearning by retraining only one shard instead of the entire model and was therefore selected as the optimal setup.

D. Main Results and Insights

The primary objective of this work was to address three key limitations observed in prior machine unlearning approaches for EEG-based seizure detection: high unlearning latency, residual interference caused by patient similarity, and the lack of support for online data removal. Existing unlearning strategies based on knowledge distillation require retraining auxiliary teacher models and performing additional student training cycles, resulting in unlearning costs that are often multiple times higher than the original training time.

First, the proposed SISA-based training framework significantly reduces unlearning time. By partitioning the dataset into independent shards and retraining only the affected shard when a patient's data must be removed, the unlearning cost is reduced to approximately $\frac{1}{n}$ of the full retraining time, where n is the number of shards. In contrast to prior distillation-based approaches that required nearly three times the original training duration, SISA enables scalable and computationally efficient unlearning suitable for real-world deployment.

Second, EEG data often exhibit strong inter-patient similarity, which can lead to residual influence when removing a specific patient's data. To address this issue, patients with similar EEG patterns were intentionally distributed across different shards using a knowledge-aware sharding strategy. This design ensures that unlearning a single patient does not suffer from representational leakage caused by correlated samples from other patients, thereby improving the reliability and verifiability of the unlearning process.

Third, Influence Function–based unlearning was employed to support online unlearning, enabling immediate removal of specific data points without retraining. By directly estimating and reversing the contribution of the forget set from the trained model parameters, this approach allows data to be removed dynamically while preserving overall detection performance. Together, the combination of SISA for efficient offline unlearning and Influence Functions for fast online unlearning provides a flexible and practical unlearning framework.

To ensure that SISA-based unlearning does not degrade seizure detection performance, an additional shard-size optimization experiment was conducted. By comparing multiple shard configurations, the results confirmed that appropriate shard granularity is essential for balancing unlearning efficiency and predictive accuracy. The two-shard configuration achieved the best trade-off, validating the effectiveness of the proposed design choices.

E. Key Observations

The experimental results highlight several important insights. First, SISA training exhibits a clear trade-off between unlearning efficiency and seizure detection performance, with shard granularity playing a critical role. Second, Influence Function–based unlearning enables fast, retraining-free deletion and can achieve precise patient-level forgetting, though its effectiveness depends on model architecture and data distribution. Finally, combining SISA-based exact unlearning with IF-based approximate unlearning provides a flexible dual-mode framework capable of supporting both offline retraining and real-time deletion requests in clinical seizure detection systems.

VII. Conclusion and Future Work

This work presented a comprehensive machine unlearning framework for EEG-based seizure detection, addressing the critical need for privacy-preserving and adaptable healthcare AI systems. Two complementary unlearning strategies were implemented and evaluated: SISA-based modular retraining,

TABLE III
IF-BASED UNLEARNING IMPACT ON MODEL PERFORMANCE

Model	Phase	Accuracy	ROC-AUC	F1-Score	Recall
EEGNet – Balanced	Before	0.517	0.812	0.618	0.971
	After	0.598	0.866	0.219	0.333
EEGNet – MaxPerformance	Before	0.925	0.979	0.905	0.886
	After	0.908	0.994	0.873	0.786
EEGNet – MaxReliability	Before	0.580	0.794	0.629	0.886
	After	0.592	0.788	0.636	0.886
TCN – Balanced	Before	0.402	0.879	0.574	1.000
	After	0.402	0.878	0.574	1.000
TCN – MaxPerformance	Before	0.460	0.502	0.447	0.543
	After	0.667	0.754	0.293	0.171
TCN – MaxReliability	Before	0.391	0.280	0.562	0.971
	After	0.402	0.280	0.574	0.980

TABLE IV
SISA SHARD EXPERIMENT RESULTS

Model	Shard Count	Patients /Shard	Accuracy	ROC -AUC
PCT_full	1 (Baseline)	12	0.904	0.495
PCT_2_shard	2	6	0.778	0.666
PCT_3_shard	3	4	0.711	0.515
PCT_4_shard	4	3	0.643	0.550

which enables exact and efficient unlearning by retraining only affected data shards, and Influence Function–based unlearning, which supports fast, retraining-free removal of specific data samples. Through extensive experimentation, including shard size analysis, the study demonstrated that a two-shard SISA configuration achieves the best balance between detection performance, computational efficiency, and unlearning cost.

A key contribution of this work is the introduction of a Knowledge-Aware Sharding Encoder, which leverages patient-level EEG similarity to improve shard formation. By distributing semantically similar patients across different shards, the framework reduces cross-patient interference and mitigates residual memory leakage during unlearning. In parallel, the IF-based unlearning pipeline was successfully applied to EEGNet and TCN architectures, demonstrating the feasibility of online unlearning without compromising overall model scalability. Collectively, the results highlight important trade-offs between accuracy, robustness, and unlearning speed, and show that combining SISA and IFs provides a flexible dual-mode unlearning solution suitable for real-world clinical deployment.

Despite these strengths, several limitations remain and motivate future research. Influence Function–based unlearning relies on second-order approximations, which may introduce approximation error in highly non-convex models. Additionally, SISA performance is sensitive to shard size, requiring careful configuration to avoid underfitting or excessive retraining overhead. Future work should explore adaptive shard sizing strategies and more robust influence estimation techniques.

Promising future directions include extending the framework to larger and more diverse EEG datasets to assess generalization across institutions and patient populations, and evaluating unlearned models on unseen subjects to better understand post-unlearning generalization. Further research may integrate unlearning with federated learning to support decentralized, privacy-aware training across multiple hospitals. Evaluating the framework against stronger privacy attacks, such as advanced membership inference and reconstruction attacks, is another important extension. Finally, enabling real-time and on-device unlearning for wearable or implantable EEG systems represents a critical step toward fully deployable, regulation-compliant seizure detection systems.

Overall, this work establishes a strong foundation for privacy-preserving healthcare AI and demonstrates that high-performance seizure detection can coexist with modern data protection requirements such as the Right to Be Forgotten.

REFERENCES

[1] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine Unlearning," *Proc. IEEE Symposium on Security and Privacy (S&P)*, pp. 141–159, 2021.

[2] J. Xu, Z. Wu, C. Wang, and X. Jia, "Machine Unlearning: Solutions and Challenges," *IEEE Computational Intelligence Magazine*, vol. 19, no. 2, pp. 45–68, 2024.

[3] M. Jegorova, C. Kaul, C. Mayor, *et al.*, "Machine Unlearning for Seizure Prediction," *IEEE Journals & Magazine*, IEEE Xplore.

[4] X. Liu, S. Zhu, and J. Zhang, "Learn to Forget: Machine Unlearning via Neuron Masking," *arXiv preprint arXiv:2003.10933*, 2020.

[5] T. Shaik, V. Reddy, and S. Ahmed, "Exploring the Landscape of Machine Unlearning: A Comprehensive Survey and Taxonomy," *arXiv preprint arXiv:2305.15011*, 2023.

[6] T. Baumhauer, S. Eger, and I. Augenstein, "Unlearning Bias in Neural Classification Models," *Proc. EMNLP*, pp. 8303–8317, 2020.

[7] C. Becker and T. Liebig, "Certified Data Removal in Sum-Product Networks," *Proc. IEEE Int. Conf. on Knowledge Graphs (ICKG)*, pp. 225–232, 2022.

[8] X. Chen, Y. Wang, J. Xu, and X. Jia, "Graph Unlearning via Knowledge Distillation," *Proc. NeurIPS*, pp. 1–14, 2022.

[9] V. S. Chundawat, S. Tople, and A. Sharma, "Zero-Shot Machine Unlearning," *Proc. ICML*, pp. 5123–5140, 2023.

[10] M. D'Angelo, A. Pietracaprina, and J. Schneider, "How to Make Reproducible Research in Machine Unlearning with ERASURE," *IJCAI Workshop on Machine Unlearning*, pp. 1–12, 2025.

[11] A. Ginart, M. Guan, G. Valiant, and J. Zou, "Making AI Forget: Data Deletion in Machine Learning," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3513–3526, 2019.

[12] L. Graves, V. Nagisetty, and A. Ganesh, "Amnesiac Machine Learning," *arXiv preprint arXiv:2010.10981*, 2020.

[13] Y. Guo, J. Liu, and T. Chen, "Certified Data Removal in Machine Learning," *Proc. KDD*, pp. 1515–1525, 2020.
[14] W. Hao, W. He, and J. Chen, "Fast Approximate Unlearning via Influence Estimation," *Proc. AAAI Conf. on Artificial Intelligence*, vol. 38, no. 5, pp. 6201–6209, 2024.
[15] Z. Izzo, C. Huang, and X. Yao, "Approximate Data Deletion from Trained Neural Networks," *arXiv preprint arXiv:2009.04899*, 2021.
[16] Z. Kurmanji, J. Hayes, E. Triantafillou, and N. Papernot, "Towards Unbounded Machine Unlearning," *arXiv preprint arXiv:2303.13517*, 2023.
[17] D. Li, Y. Wang, and C. Wang, "Unlearning with Knowledge Distillation in Deep Neural Networks," *Pattern Recognition*, vol. 152, p. 109843, 2024.
[18] J. Marchant, R. T. Q. Chen, and R. Bunel, "Hard to Forget: Poisoning Attacks on Certified Machine Unlearning," *Proc. AAAI*, vol. 36, no. 8, pp. 8427–8435, 2022.
[19] T. T. Nguyen, L. Trinh, C. Xiao, *et al.*, "A Survey of Machine Unlearning," *arXiv preprint arXiv:2209.02299*, 2022.
[20] M. Pawelczyk, M. Schäfer, M. Asgari, and G. Kasneci, "Machine Unlearning Fails to Remove Data Poisoning Attacks," *arXiv preprint arXiv:2406.00096*, 2024.
[21] A. Sekhari, B. Ustun, and C. Zhang, "Remember What You Want to Forget: Algorithms for Machine Unlearning," *arXiv preprint arXiv:2103.03279*, 2021.
[22] A. Tarun, A. Sharma, and S. Tople, "Fast Machine Unlearning Without Retraining," *Proc. USENIX Security Symposium*, pp. 2301–2318, 2023.
[23] A. Thudi, B. Jayaraman, and D. Evans, "On the Robustness of Machine Unlearning," *Proc. IEEE Security and Privacy Workshops*, pp. 1–8, 2022.
[24] A. Wang, N. Papernot, and B. Zhang, "KGA: A General Machine Unlearning Framework Based on Knowledge Gap Alignment," *ACL Workshop on Trustworthy Machine Learning*, pp. 12–24, 2023.
[25] X. Wu, R. He, and D. Tao, "FedEraser: Enabling Efficient Client-Side Machine Unlearning in Federated Learning," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2022.
[26] Y. Cao and J. Yang, "Towards Making Systems Forget with Machine Unlearning," *Proc. IEEE Symposium on Security and Privacy (S&P)*, pp. 463–480, 2015.