End-to-End Diffusion Modeling for Clinical EEG Abnormality Detection with Spatial Filtering and Attention

Anonymous Authors[1]

Abstract

Automated interpretation of clinical electroencephalography (EEG) is challenging due to signal heterogeneity, noise contamination, and inter-subject variability. We propose DiffSA-EEG, a diffusion-based EEG classification framework that integrates learnable spatial filtering, stacked denoising autoencoders (SDA), and convolutional block attention modules (CBAM) within an end-to-end discriminative pipeline. Unlike prior diffusion-based EEG studies that focus on data generation or augmentation, our framework leverages denoising diffusion probabilistic models (DDPMs) directly for discriminative classification as a feature regularizer. We evaluate DiffSA-EEG on two large-scale clinical EEG corpora: the Temple University Hospital Abnormal EEG Corpus (TUAB) and the Temple University Epilepsy Corpus (TUEP). DiffSA-EEG consistently outperforms established baselines—including EEGNet, Deep4Net, ChronoNet, temporal convolutional networks, and the Diff-E backbone—across accuracy, AUC-ROC, and AUC-PR. Ablation analyses reveal that optimal component combinations are dataset-dependent: spatial filtering with SDA is most effective for TUAB, while SDA with CBAM yields superior performance on TUEP. Grad-CAM-based interpretability analysis further shows that the model captures clinically plausible spatial patterns aligned with established neurophysiological biomarkers.

1. Introduction

Electroencephalography (EEG) is a cornerstone non-invasive neuroimaging technique widely used for diagnosing neurological disorders such as epilepsy and stroke (Acharya et al., 2012; Obeid & Picone, 2016). However, clinical EEG interpretation remains heavily reliant on expert visual inspection, which is subjective and labor-intensive, motivating the development of automated deep learning approaches (Roy et al., 2019b).

Recent deep learning methods—including EEGNet (Lawhern et al., 2018), Deep4Net (Schirrmeister et al., 2017), ChronoNet (Roy et al., 2019a), and temporal convolutional networks (TCN) (Gemein et al., 2020)—have shown promising results on benchmark datasets such as TUAB and TUEP. More recently, transformer-based and state-space models have further advanced performance by capturing long-range temporal dependencies (Tegon et al., 2025). In parallel, denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) have emerged as a powerful tool in biomedical signal processing. For example, Diff-E demonstrated the potential of conditional diffusion models for decoding imagined speech from EEG (Kim et al., 2023), while DiffMDD applied diffusion-based learning for major depressive disorder diagnosis (Wang et al., 2024). However, existing diffusion-based EEG studies have primarily focused on data generation or augmentation (Shu et al., 2025; Jiang et al., 2024), rather than leveraging diffusion as a core component for end-to-end discriminative classification on large-scale clinical datasets.

To address this gap, we propose **DiffSA-EEG**, a diffusion-based EEG classification framework that integrates: (1) a learnable *spatial filter* (SF) for topographical dimensionality reduction, (2) a *stacked denoising autoencoder* (SDA) for noise-invariant latent representations, and (3) *convolutional block attention modules* (CBAM) (Woo et al., 2018) for selective feature emphasis. Unlike prior work, DiffSA-EEG uses diffusion modeling as a trajectory-based regularization mechanism within a unified encoder–decoder architecture, jointly optimized with a conditional autoencoder (CAE) and classification head.

Our key contributions are: (i) a novel integration of diffusion modeling into discriminative EEG classification, (ii) a systematic ablation study revealing dataset-dependent optimal configurations, and (iii) Grad-CAM-based interpretability analysis demonstrating clinically plausible spatial patterns.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Methods

2.1. Framework Overview

DiffSA-EEG extends the Diff-E architecture (Kim et al., 2023) with a unified encoder–decoder framework comprising four core modules (Figure 1):

(1) DDPM Backbone. Given a clean EEG signal $X_0 \in \mathbb{R}^{C\times T}$, the forward diffusion process produces a noisy sample $X_t = \sqrt{\bar{\alpha}_t}X_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ is the cumulative signal retention ratio. The reverse process, parameterized by a time-conditional U-Net (Ronneberger et al., 2015), directly reconstructs X_0 from X_t, providing signal-level guidance for downstream tasks. Unlike conventional diffusion studies that focus on noise prediction, our model directly predicts the clean signal, enabling the diffusion module to serve as a feature regularizer. The diffusion objective minimizes $\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t,\epsilon}[\|X_0 - f_\theta(X_t, t)\|_2^2]$.

(2) Learnable Spatial Filter (SF). To reduce channel redundancy, an SVD-based spatial filter computes $F = W_f U^\top X_0 \in \mathbb{R}^{M\times T}$, where $U \in \mathbb{R}^{C\times M}$ contains the dominant M spatial bases and $W_f \in \mathbb{R}^{M\times M}$ is a learnable projection matrix. An inverse spatial filter recovers refined representations after diffusion-guided reconstruction.

(3) Stacked Denoising Autoencoder (SDA). Following Vincent et al. (2010), Gaussian noise is injected into latent features ($\tilde{z}^{(i)} = z^{(i)} + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, $\sigma = 0.05$) through three stacked convolutional layers, encouraging noise-invariant representations shared by both the decoder and the classification head.

(4) CBAM. Channel and temporal attention mechanisms (Woo et al., 2018) are sequentially applied after residual convolution blocks to emphasize informative EEG features. Self-attention blocks (Vaswani et al., 2017) are additionally incorporated in deeper layers to capture long-range temporal dependencies.

Conditional Autoencoder (CAE) corrects diffusion-induced reconstruction errors via skip connections from intermediate DDPM layers, trained with a Smooth L1 loss. Both the original input X_0 and the DDPM output $\hat{X}_0$ are injected into the final decoder layer, enabling the CAE to leverage both global structure and local refinement cues. The overall objective jointly optimizes all components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DDPM}} + \mathcal{L}_{\text{CAE}} + \mathcal{L}_{\text{CLS}} \tag{1}$$

2.2. Datasets and Preprocessing

TUAB (v3.0.1) comprises 2,993 EEG recordings (1,521 normal, 1,472 abnormal) from 2,130 patients diagnosed with conditions such as epilepsy, stroke, depression, and Alzheimer's disease. **TUEP** (v2.0.1) contains 2,298 recordings from 200 patients (513 non-epileptic, 1,785 epileptic); after exclusion criteria (duration $<$ 130s or missing channels), 317 non-epileptic and 1,275 epileptic recordings remained, with k-means SMOTE (Last et al., 2018) applied for class balancing. Both datasets used 21 channels (international 10–20 system), downsampled to 250 Hz, with 1 Hz high-pass Butterworth and 60 Hz notch filters applied. The first 60 seconds were segmented with time-reversal augmentation (Rommel et al., 2022), yielding two non-overlapping segments per recording, normalized via z-scoring.

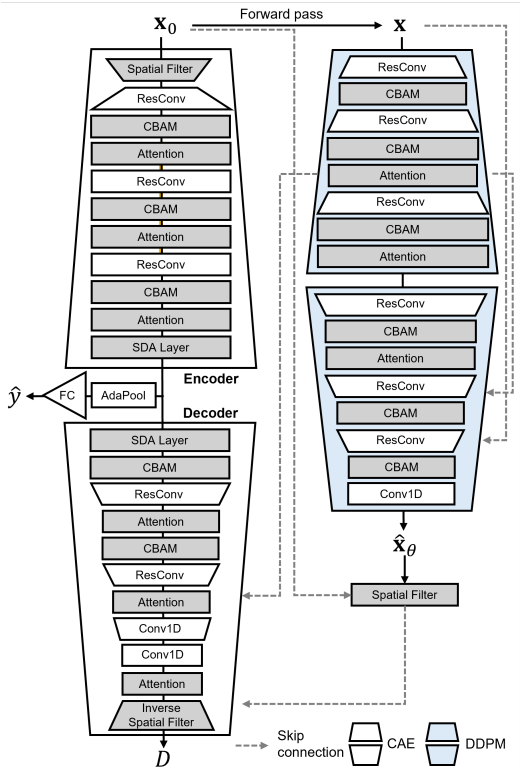

Figure 1. Overview of DiffSA-EEG. The DDPM backbone generates diffusion-refined representations, which are corrected by the CAE and classified via adaptive pooling and a fully connected head. SF, SDA, and CBAM modules are integrated throughout the encoder–decoder pathway.

3. Experiments and Results

All models were trained for 100 epochs across 10 independent runs with different random seeds on an NVIDIA RTX A6000 GPU. For TUAB, official train/test splits were used; for TUEP, 80/20 random splits were repeated 10 times. Statistical significance was assessed via bootstrap-based two-sample t-tests (10,000 iterations). All results are reported as mean $\pm$ standard error (SE).

Table 1. Performance comparison on TUAB and TUEP (mean ± SE). Best results in **bold**. All improvements of DiffSA-EEG over baselines are statistically significant ($p < 0.05$).

Model	TUAB					
	Acc (%)	Recall (%)	Spec (%)	AUC-ROC (%)	AUC-PR (%)	Bal Acc (%)
EEGNet	78.0±0.4	77.7±0.4	74.4±1.4	83.3±0.3	55.0±5.0	46.8±5.0
Deep4Net	82.5±0.6	82.0±0.5	76.0±1.1	89.5±0.4	90.0±0.4	81.9±0.5
ChronoNet	81.7±0.7	81.2±0.7	75.0±0.8	87.8±0.8	88.4±0.8	81.2±0.8
TCN	79.9±0.1	79.3±0.2	72.1±1.1	84.7±0.2	84.7±0.1	79.3±0.2
Diff-E	82.6±0.1	81.8±0.1	72.1±0.8	88.2±0.1	87.6±0.2	76.9±0.5
DiffSA-EEG	**83.9±0.3**	**83.6±0.3**	**79.4±0.8**	**90.6±0.3**	**89.4±0.5**	**81.5±0.6**

Model	TUEP					
	Acc (%)	Recall (%)	Spec (%)	AUC-ROC (%)	AUC-PR (%)	Bal Acc (%)
EEGNet	58.1±1.8	58.3±1.8	17.6±3.8	90.0±1.1	54.5±2.9	46.8±5.0
Deep4Net	85.2±1.1	85.3±1.1	71.2±2.3	96.7±0.3	94.7±0.9	81.9±0.5
ChronoNet	90.4±0.3	90.4±0.3	88.2±1.1	96.5±0.2	96.0±0.3	81.2±0.8
TCN	91.6±0.1	91.6±0.1	87.9±0.4	96.5±0.1	93.8±0.3	79.3±0.2
Diff-E	94.6±0.1	94.6±0.1	92.0±0.5	98.1±0.1	97.6±0.1	93.3±0.3
DiffSA-EEG	**94.7±0.2**	**94.7±0.2**	**93.2±0.2**	**98.3±0.1**	**97.8±0.3**	**94.0±0.2**

3.1. Comparison with Baselines

Table 1 presents the comprehensive comparison between DiffSA-EEG and five baseline models across six evaluation metrics. On TUAB, DiffSA-EEG achieved 83.9% accuracy and 90.6% AUC-ROC, outperforming all baselines with statistically significant margins ($p < 0.05$). The improvements were particularly notable in specificity (79.4%) and AUC-PR (89.4%), demonstrating the model's robustness against false-positive predictions. On TUEP, DiffSA-EEG achieved 94.7% accuracy, 93.2% specificity, and 98.3% AUC-ROC. While Diff-E performed competitively on TUEP (94.6% accuracy), DiffSA-EEG showed consistent advantages across all metrics, with the largest improvements in specificity and balanced accuracy. Notably, EEGNet exhibited substantial performance degradation on TUEP (58.1% accuracy), likely due to the dataset's class imbalance and complex epileptiform patterns, whereas DiffSA-EEG maintained robust performance.

3.2. Ablation Study

We systematically evaluated all $2^4 = 16$ combinations of the four modules (SF, SDA, CBAM, self-attention) added to the Diff-E backbone (Table 2). A key finding is the *non-monotonic relationship* between architectural complexity and performance: task-specific two-component configurations consistently outperformed both simpler and fully integrated models.

On TUAB, the **SF + SDA** combination achieved the best results (84.6% accuracy, 90.8% AUC-ROC), outperforming even the full four-module configuration (83.9% accuracy). This indicates that spatial dimensionality reduction and noise robustness are most critical for heterogeneous clinical abnormalities, while the addition of attention mechanisms introduces unnecessary redundancy. On TUEP, the **SDA + CBAM** configuration yielded superior performance (96.0% accuracy, 98.4% AUC-ROC), surpassing the full model by 1.3% in accuracy. This suggests that attention-guided refinement is more effective for capturing the transient, spatially structured epileptiform discharges characteristic of TUEP.

3.3. Interpretability via Grad-CAM

To assess clinical plausibility, we conducted a Grad-CAM-based relevance analysis (Selvaraju et al., 2017) by projecting component-level importance back to the original scalp montage via the SVD spatial bases used during the forward spatial filtering process (Figure 2). For TUAB, the model attributed higher importance to centro-posterior and temporo-occipital regions for abnormal EEG detection, consistent with posterior slowing patterns characteristic of diffuse cerebral dysfunction. For TUEP, increased rele-

Table 2. Top ablation configurations on TUAB and TUEP (mean ± SE). Full model includes all four modules.

Configuration	TUAB			TUEP		
	Acc (%)	AUC-ROC (%)	Bal Acc (%)	Acc (%)	AUC-ROC (%)	Bal Acc (%)
Diff-E (base)	82.6±0.1	88.2±0.1	76.9±0.5	94.6±0.1	98.1±0.1	93.3±0.3
SF + SDA	**84.6±0.1**	**90.8±0.2**	**81.6±0.6**	94.9±0.2	98.5±0.1	94.6±0.3
SDA + CBAM	82.1±0.2	87.9±0.3	79.7±0.8	**96.0±0.2**	98.4±0.2	**95.1±0.2**
SF + SDA + CBAM	84.2±0.3	90.6±0.3	82.6±0.6	95.4±0.2	**98.6±0.1**	94.8±0.3
Full (all four)	83.9±0.3	90.6±0.3	81.5±0.6	94.7±0.2	98.3±0.1	94.0±0.2

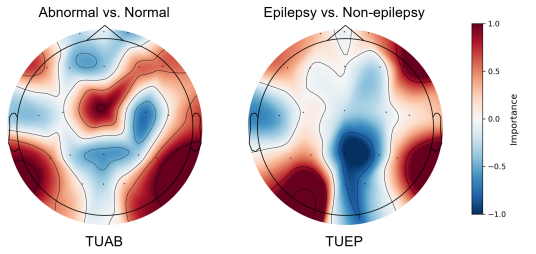

Figure 2. Average difference in Grad-CAM scalp distributions. Left: abnormal vs. normal (TUAB). Right: epilepsy vs. non-epilepsy (TUEP). Positive values indicate higher model-attributed importance for the target class.

vance was observed in right temporal and inferior-lateral regions, aligning with the well-documented temporal lobe involvement in epileptic activity. These cohort-specific spatial patterns demonstrate that DiffSA-EEG learns clinically interpretable representations rather than relying on a single shared topographic signature.

4. Discussion and Conclusion

We presented DiffSA-EEG, a diffusion-based framework for clinical EEG classification that integrates spatial filtering, denoising, and attention mechanisms within an end-to-end pipeline. Our work offers three key insights for the structured health data community.

First, *diffusion as regularization*: unlike conventional approaches that use DDPMs for data generation, DiffSA-EEG employs diffusion as a trajectory-based regularizer, enforcing consistency across progressively perturbed signal realizations. This is particularly beneficial for clinical EEG, which is often contaminated with heterogeneous artifacts and non-stationary noise.

Second, *task-specific modularity*: the non-monotonic relationship between model complexity and performance underscores the importance of selective component integration over monolithic design. The TUAB dataset benefits primarily from spatial dimensionality reduction (SF) and artifact suppression (SDA), while TUEP responds more favorably to attention-based feature refinement (CBAM). This has practical implications for deploying EEG models in resource-limited clinical environments.

Third, *clinical interpretability*: the Grad-CAM analysis reveals that the learned representations align with established neurophysiological biomarkers—posterior slowing for diffuse cerebral dysfunction and temporal lobe involvement for epileptic activity—supporting the model's potential as an explainable diagnostic aid.

Limitations include the need for multi-site validation to fully assess clinical readiness, and the computational overhead of diffusion-based processing. Future work will explore lightweight diffusion variants for real-time deployment and extend the framework to additional structured clinical time-series modalities.

References

Acharya, U. R., Molinari, F., Sree, S. V., Chattopadhyay, S., Ng, K.-H., and Suri, J. S. Automated diagnosis of epileptic EEG using entropies. *Biomedical Signal Processing and Control*, 7(4):401–408, 2012.

Gemein, L. A., Schirrmeister, R. T., Chrabaszcz, P., Wilson, D., Boedecker, J., Schulze-Bonhage, A., Hutter, F., and Ball, T. Machine-learning-based diagnostics of EEG pathology. *NeuroImage*, 220:117021, 2020.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020.

Jiang, Z., Dai, W., Wei, Q., Qin, Z., Li, K., and Zhang, L. EEG-DIF: Early warning of epileptic seizures through generative diffusion model-based multi-channel EEG signals forecasting. *arXiv preprint arXiv:2410.17343*, 2024.

Kim, S., Lee, Y.-E., Lee, S.-H., and Lee, S.-W. Diff-E: Diffusion-based learning for decoding imagined speech EEG. *arXiv preprint arXiv:2307.14389*, 2023.

Last, F., Douzas, G., and Bacao, F. Oversampling for imbal-

anced learning based on K-means and SMOTE. *Information Sciences*, 465:1–20, 2018.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15 (5):056013, 2018.

Obeid, I. and Picone, J. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016.

Rommel, C., Moreau, T., Paillard, J., and Gramfort, A. CADDA: Class-wise automatic differentiable data augmentation for EEG signals. *arXiv preprint arXiv:2106.13695*, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Roy, S., Kiral-Kornek, I., and Harrer, S. ChronoNet: A deep recurrent neural network for abnormal EEG identification. In *Artificial Intelligence in Medicine*, pp. 47–56. Springer, 2019a.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. Deep learning-based electroencephalography analysis: a systematic review. volume 16, pp. 051001. IOP Publishing, 2019b.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., and Ball, T. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Shu, K., Wu, L., Zhao, Y., Liu, A., Qian, R., and Chen, X. Data augmentation for seizure prediction with generative diffusion model. *IEEE Transactions on Cognitive and Developmental Systems*, 17(3):577–591, 2025.

Tegon, A., Ingolfsson, T. M., Wang, X., Benini, L., and Li, Y. FEMBA: Efficient and scalable EEG analysis with a bidirectional Mamba foundation model. *arXiv preprint arXiv:2502.06438*, 2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

Wang, Y. et al. DiffMDD: A diffusion-based deep learning framework for MDD diagnosis using EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:728–738, 2024.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. CBAM: Convolutional block attention module. In *European Conference on Computer Vision*, pp. 3–19. Springer, 2018.