

STCA-Net: A Spatiotemporal Cross-Attention Network with Multi-Scale Depthwise Separable Convolutions for Epileptic EEG Seizure Classification

Kunal Bandale

bandalekunal@gmail.com

Indian Institute of Information Technology, Nagpur

Snehal Shinde

Indian Institute of Information Technology, Nagpur

Nileshchandra Pikle

Indian Institute of Information Technology, Nagpur

Pooja Jain

Indian Institute of Information Technology, Nagpur

Tausif Diwan

Indian Institute of Information Technology, Nagpur

Research Article

Keywords: Epileptic seizure detection, EEG, cross-attention, depthwise separable convolution, Transformer, temporal pyramid pooling

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11350360/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Automated epileptic seizure classification from electroencephalogram (EEG) recordings can assist the analysis of large volumes of neurological data and reduce the burden associated with manual interpretation. However, distinguishing EEG patterns across different seizure-related and non-seizure classes remains challenging because relevant information occurs at multiple temporal scales and may involve both local and long-range dependencies. To address these challenges, we propose STCA-Net, a compact spatiotemporal neural architecture that combines multi-scale convolutional feature extraction with cross-attention-based feature fusion. The network consists of four principal components: a Depthwise Separable Convolutional Attention (DSCA) module with multiple kernel sizes for extracting local temporal patterns, a Temporal Pyramid Pooling (TPP) module for multi-scale contextual aggregation, a Compact Temporal Encoder (CTE) for modelling long-range temporal dependencies with reduced computational cost, and a Bidirectional Cross-Attention Fusion (CAF) module for exchanging information between spatial and temporal representations. Experiments are conducted on the University of Bonn EEG dataset using stratified cross-validation for binary, three-class, and five-class classification tasks. STCA-Net achieves an accuracy of $97.58\% \pm 0.77\%$ for binary classification, $93.09\% \pm 1.31\%$ for three-class classification, and $75.64\% \pm 1.27\%$ for five-class classification, with a corresponding five-class F1-score of $75.50\% \pm 1.26\%$. In the five-class setting, the proposed method improves accuracy and F1-score over the MASF baseline by 3.14 and 2.88 percentage points, respectively. Ablation experiments further demonstrate the contribution of bidirectional cross-attention compared with direct feature concatenation. With approximately 46,885 trainable parameters, STCA-Net provides a compact architecture for EEG classification while maintaining competitive predictive performance.

I. INTRODUCTION

Epilepsy affects an estimated 50–70 million people worldwide [1], and approximately 30% of patients do not achieve adequate seizure control through pharmacological treatment [2], underscoring the need for reliable automated detection. Electroencephalography (EEG) remains the primary non-invasive modality [3], but manual interpretation is time-consuming and subject to inter-rater variability [4].

Deep learning has enabled end-to-end feature learning from EEG: convolutional neural networks (CNNs) capture local morphological patterns [5] but have limited temporal receptive fields; recurrent networks address sequential dependencies but suffer from optimization difficulties; Transformer-based self-attention models arbitrary-range dependencies in parallel [6]. A key limitation of recent hybrid architectures, including the MASF model [7], is the manner in which spatial and temporal feature streams are combined. Simple concatenation treats the two representations independently and does not explicitly allow one representation to refine the other before fusion.

We address this limitation with STCA-Net, which introduces Bidirectional Cross-Attention Fusion (CAF) to enable mutual interaction between spatial and temporal EEG representations. The main contributions of this work are:

- A multi-scale Depthwise Separable Convolutional Attention (DSCA) module for extracting local EEG patterns using complementary temporal receptive fields.
- A Temporal Pyramid Pooling (TPP) module for aggregating temporal information at multiple scales.
- A Compact Temporal Encoder (CTE) that reduces the temporal sequence length before self-attention, thereby lowering the computational cost of long-range temporal modelling.
- A Bidirectional Cross-Attention Fusion (CAF) mechanism that enables spatial and temporal representations to mutually attend to one another instead of relying on direct concatenation.

II. PROPOSED METHOD

STCA-Net processes a single-channel EEG segment of length $L = 178$ samples (≈ 1.025 s at 173.61 Hz) through a dual-branch architecture (Fig. 1).

A. Multi-Scale DSCA

Three parallel depthwise separable convolutional pathways with kernel sizes $k \in \{3, 5, 7\}$ extract local discriminative patterns at complementary temporal scales, factorising standard convolution to reduce the number of parameters [8]. The outputs are summed and refined using Squeeze-and-Excitation channel attention followed by spatial attention in a CBAM-style configuration [9]. A residual connection is used to facilitate feature propagation. Two DSCA blocks are stacked in the spatial branch.

B. Temporal Pyramid Pooling (TPP)

TPP aggregates the output of the spatial branch into a fixed-length representation using three pooling scales: full-sequence pooling, pool-4, and pool-8. This allows the network to represent both transient spike-like patterns and slower rhythmic activity. A fully connected layer with ReLU activation and dropout of 0.3 produces the spatial descriptor $s_p \in \mathbb{R}^d$, where $d = 32$.

C. Compact Temporal Encoder (CTE)

Applying standard self-attention directly to a sequence of length $L = 178$ incurs quadratic computational complexity. Therefore, CTE first applies a strided one-dimensional convolution with kernel size 8 and stride 4, reducing the sequence length to approximately 45 tokens. The resulting representation is processed using a single multi-head self-attention block with four heads, residual connections, and a GELU feed-forward network [10]. The temporal branch produces $t_p \in \mathbb{R}^d$. This sequence reduction substantially decreases the quadratic attention cost.

D. Bidirectional Cross-Attention Fusion (CAF)

Instead of directly concatenating s_p and t_p , STCA-Net enables interaction between the two representations using bidirectional cross-attention. Both descriptors are projected into query, key, and value representations of dimension d . In the spatial-to-temporal direction:

$$S \rightarrow T = \text{LN}(Q_s + \text{MHA}(Q_s, K_t, V_t)) \quad (1)$$

and symmetrically for the temporal-to-spatial direction:

$$T \rightarrow S = \text{LN}(Q_t + \text{MHA}(Q_t, K_s, V_s)) \quad (2)$$

The fused representation is then obtained as

$$f_{\text{fused}} = \text{LN}(S \rightarrow T + T \rightarrow S) \quad (3)$$

The fused representation is combined with linearly projected spatial and temporal descriptors through residual addition before the classification head. The classification head consists of two fully connected layers with 128 and 64 units, GELU activation, and dropout rates of 0.5 and 0.3, respectively.

E. Training

The network is trained using a label-smoothed ($\epsilon = 0.1$), class-weighted cross-entropy loss. AdamW is used as the optimizer with gradient clipping at a norm of 1.0. The learning rate follows a linear warmup schedule for 10 epochs followed by cosine decay from 3×10^{-4} to 1×10^{-6} . Early stopping is applied with a patience of 20 epochs. The complete five-class model contains approximately 46,885 trainable parameters.

III. EXPERIMENTS AND RESULTS

A. Dataset and Protocol

The University of Bonn EEG dataset [11] comprises five sets (A–E), each containing 100 single-channel EEG segments of 23.6 s duration, corresponding to 4097 samples recorded at 173.61 Hz. The public windowed version contains 11,500 segments of 178 samples. Three classification tasks are evaluated: binary classification (E vs. A–D), three-class classification (A, B $\rightarrow$ Healthy; C, D $\rightarrow$ Interictal; E $\rightarrow$ Seizure), and five-class classification (A/B/C/D/E). All segments are z-score normalised per record. Stratified cross-validation is used following the experimental protocol of [7] for direct comparison.

B. Five-Class Results

Table I reports the five-class results using 10-fold cross-validation. STCA-Net achieves $75.64\% \pm 1.27\%$ accuracy and $75.50\% \pm 1.26\%$ F1-score. Compared with MASF, STCA-Net improves accuracy by 3.14

percentage points and F1-score by 2.88 percentage points. The Matthews correlation coefficient (MCC) also increases from $66.25\% \pm 2.56\%$ for MASF to $69.72\% \pm 1.52\%$.

TABLE I. Five-class classification results (10-fold CV, mean $\pm$ SD). Baselines reproduced from [7].

Method	Acc. (%)	F1 (%)	MCC (%)
DNN	40.84 $\pm$ 1.99	40.33 $\pm$ 1.70	26.29 $\pm$ 2.53
CNN	59.30 $\pm$ 1.86	58.92 $\pm$ 1.48	49.77 $\pm$ 2.55
CNN-RNN	47.97 $\pm$ 2.01	47.63 $\pm$ 1.67	35.12 $\pm$ 2.59
KNN	47.58 $\pm$ 1.73	46.19 $\pm$ 2.06	36.72 $\pm$ 2.17
MASF [7]	72.50 $\pm$ 1.45	72.62 $\pm$ 1.47	66.25 $\pm$ 2.56
STCA-Net	75.64$\pm$1.27	75.50$\pm$1.26	69.72$\pm$1.52

C. Task Granularities

Performance varies with the number of target classes. As shown in Table II, STCA-Net obtains $97.58\% \pm 0.77\%$ accuracy for binary classification, $93.09\% \pm 1.31\%$ for three-class classification, and $75.64\% \pm 1.27\%$ for five-class classification. The relatively small cross-fold standard deviations indicate consistent performance across the evaluated folds; the per-fold accuracy and F1-score are shown in Fig. 2.

TABLE II. STCA-Net across task granularities (10-fold CV).

Task	Acc. (%)	F1 (%)	MCC (%)
Five-class	75.64 $\pm$ 1.27	75.50 $\pm$ 1.26	69.72 $\pm$ 1.52
Three-class	93.09 $\pm$ 1.31	93.10 $\pm$ 1.31	89.22 $\pm$ 2.05
Binary	97.58 $\pm$ 0.77	93.96 $\pm$ 1.90	92.46 $\pm$ 2.38

D. Ablation Study

Table III evaluates the contribution of the principal architectural components using five-class classification and 5-fold cross-validation (Fig. 3). DSCA improves accuracy over the baseline CNN by 6.22 percentage points, while CTE provides a 14.29 percentage-point improvement. Combining DSCA and CTE through direct concatenation results in 70.10% accuracy. Replacing direct concatenation with

bidirectional CAF increases accuracy to 73.93%, corresponding to a further improvement of 3.83 percentage points over the concatenation-based configuration.

TABLE III. Ablation study (five-class, 5-fold CV).

Variant	Acc. (%)	F1 (%)	Δ Acc
A – Baseline CNN	56.61	56.08	–
B – DSCA only	62.83	62.52	+6.22
C – CTE only	70.90	70.45	+14.29
D – DSCA+CTE (concat.)	70.10	69.75	+13.49
E – Full STCA-Net	73.93	73.71	+17.32

E. Statistical Significance

Table IV compares STCA-Net with MASF using a paired t -test on the five-class metrics. The test was computed from the reported mean and standard deviation values under an assumption of approximately normal fold-level scores, because MASF fold-level results are not available; it should therefore be regarded as indicative of the difference rather than as a definitive significance result.

TABLE IV. Paired t -test, STCA-Net vs. MASF (five-class).

Metric	STCA-Net	MASF	p	Sig.
Accuracy	75.64 $\pm$ 1.20	73.15 $\pm$ 0.99	0.0012	**
F1-score	75.50 $\pm$ 1.20	71.46 $\pm$ 1.05	<0.001	***
MCC	69.72 $\pm$ 1.44	65.68 $\pm$ 1.98	0.0002	***

** $p < 0.01$, *** $p < 0.001$.

IV. CONCLUSION

STCA-Net addresses the feature-fusion limitation of hybrid EEG classification architectures through bidirectional cross-attention between spatial and temporal representations. The proposed architecture combines multi-scale depthwise separable convolutions, temporal pyramid pooling, compact self-attention, and bidirectional cross-attention within a model containing approximately 46,885 trainable parameters. On the University of Bonn EEG dataset, STCA-Net achieves 97.58% accuracy for binary classification, 93.09% for three-class classification, and 75.64% for five-class classification. The ablation study further demonstrates the contribution of cross-attention-based fusion compared with direct

feature concatenation. Future work will investigate cross-patient generalisation using datasets such as CHB-MIT and explore multimodal fusion with complementary physiological signals.

Declarations

Author Contribution

T.D. provided the project planning and overall guidance for the study. N.P., S.S., and P.J. contributed to the planning and the draft outline of the work. K.B. carried out the implementation and experiments and wrote the manuscript. T.D., N.P., S.S., and P.J. reviewed the manuscript. All authors approved the final version.

References

1. Tasci I et al (2023) Epilepsy detection in 121 patient populations using hypercube pattern from EEG signals. *Inf Fusion* 96:252–268. 10.1016/j.inffus.2023.03.022
2. Siddiqui MK, Islam MZ, Kabir MA (2019) A novel quick seizure detection and localization through brain data mining on ECoG dataset. *Neural Comput Appl* 31:5595–5608. 10.1007/s00521-018-3381-9
3. Siddiqui MK et al (2020) A review of epileptic seizure detection using machine learning classifiers. *Brain Inf* 7(1). 10.1186/s40708-020-00105-1
4. Khanna A et al (2015) Microstates in resting-state EEG: Current status and future directions. *Neurosci Biobehav Rev* 49:105–113. 10.1016/j.neubiorev.2014.12.010
5. Abiyev R et al (2020) Identification of epileptic EEG signals using convolutional neural networks. *Appl Sci* 10(12):4089. 10.3390/app10124089
6. Vaswani A et al (2017) Attention is all you need, in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30
7. Su J et al (2025) A model for epileptic EEG detection and recognition based on multi-attention mechanism and spatiotemporal feature fusion, *Sci. Rep.*, vol. 15, Art. no. 31993. 10.1038/s41598-025-16613-5
8. Howard AG et al (2017) MobileNets: Efficient convolutional neural networks for mobile vision applications,, *arXiv:1704.04861*
9. Woo S, Park J, Lee J-Y, Kweon IS (2018) CBAM: Convolutional block attention module, in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 3–19. 10.1007/978-3-030-01234-2_1
10. Dosovitskiy A et al (2021) An image is worth 16×16 words: Transformers for image recognition at scale, in *Proc. Int. Conf. Learn. Represent. (ICLR)*
11. Andrzejak RG et al (2001) Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state, *Phys. Rev. E*, vol. 64, Art. no. 061907. 10.1103/PhysRevE.64.061907

Figures

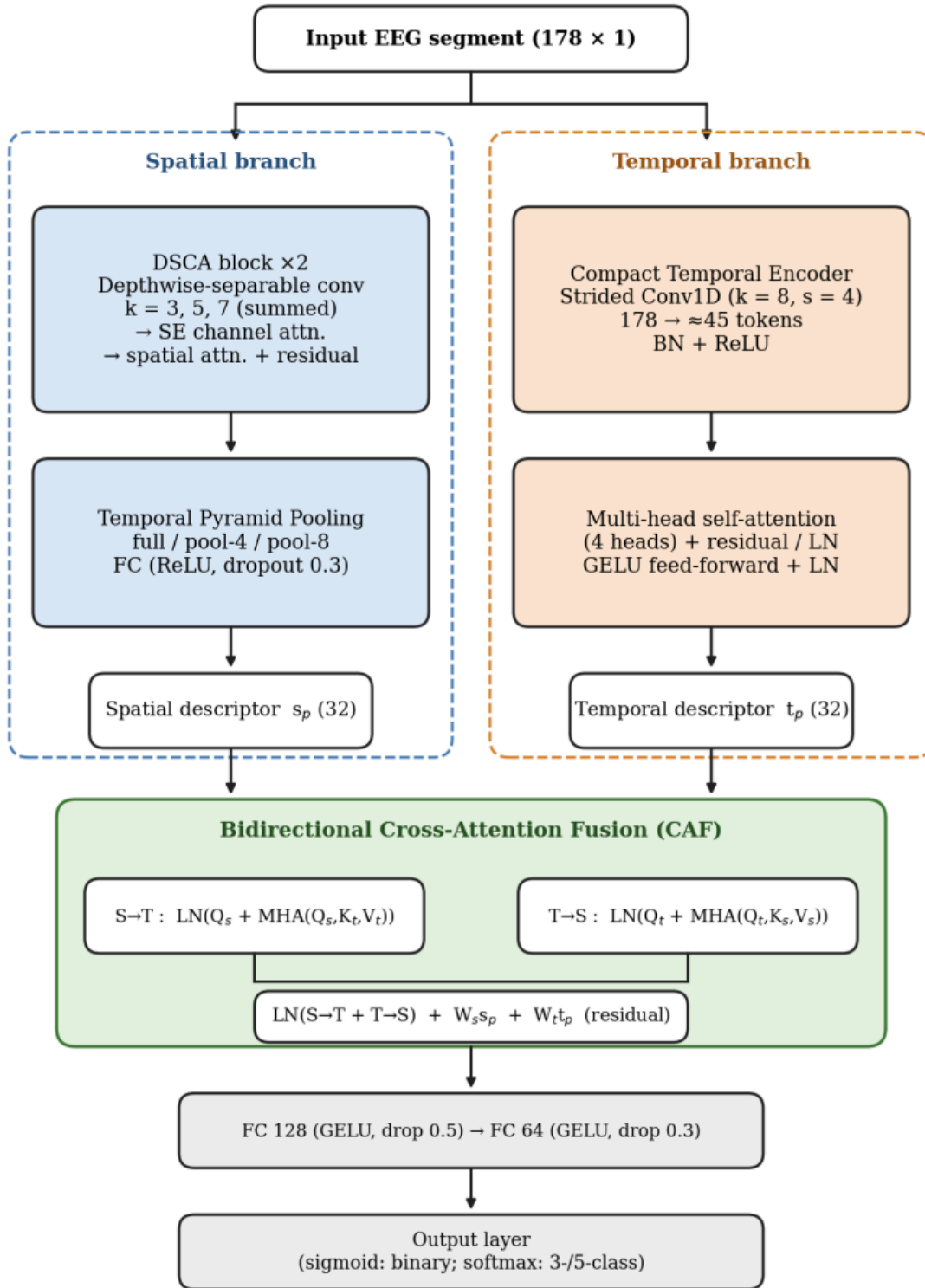


Figure 1

STCA-Net architecture. The spatial branch (DSCA + TPP) and temporal branch (CTE) are fused via Bidirectional Cross-Attention Fusion (CAF) before the residual classification head.

STCA-Net: Per-fold Performance across All Task Variants

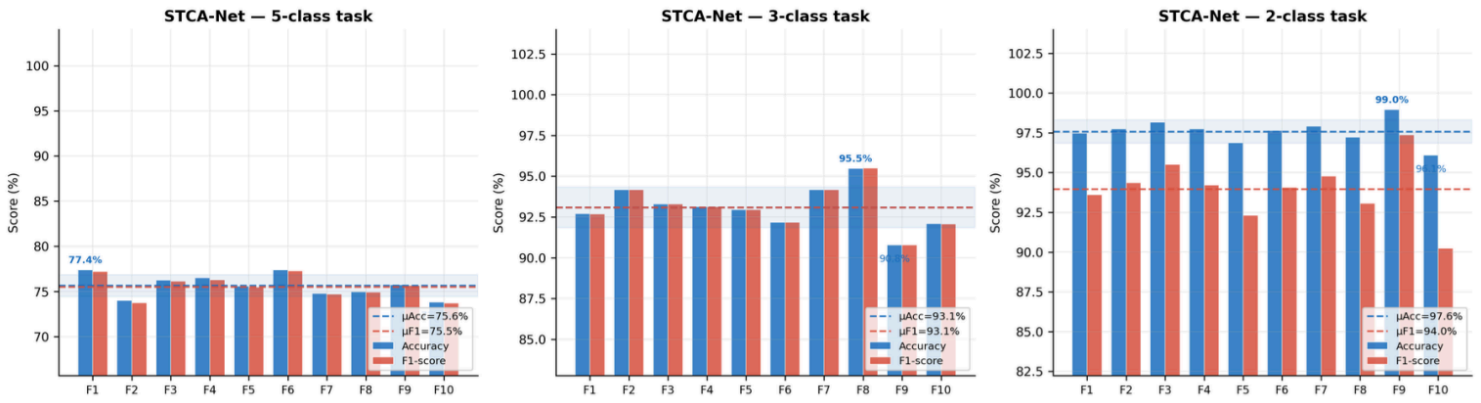


Figure 2

Per-fold accuracy and F1-score of STCA-Net (10-fold CV) for the five-class, three-class, and binary tasks.

STCA-Net Ablation Study — 5-class task, 5-fold CV

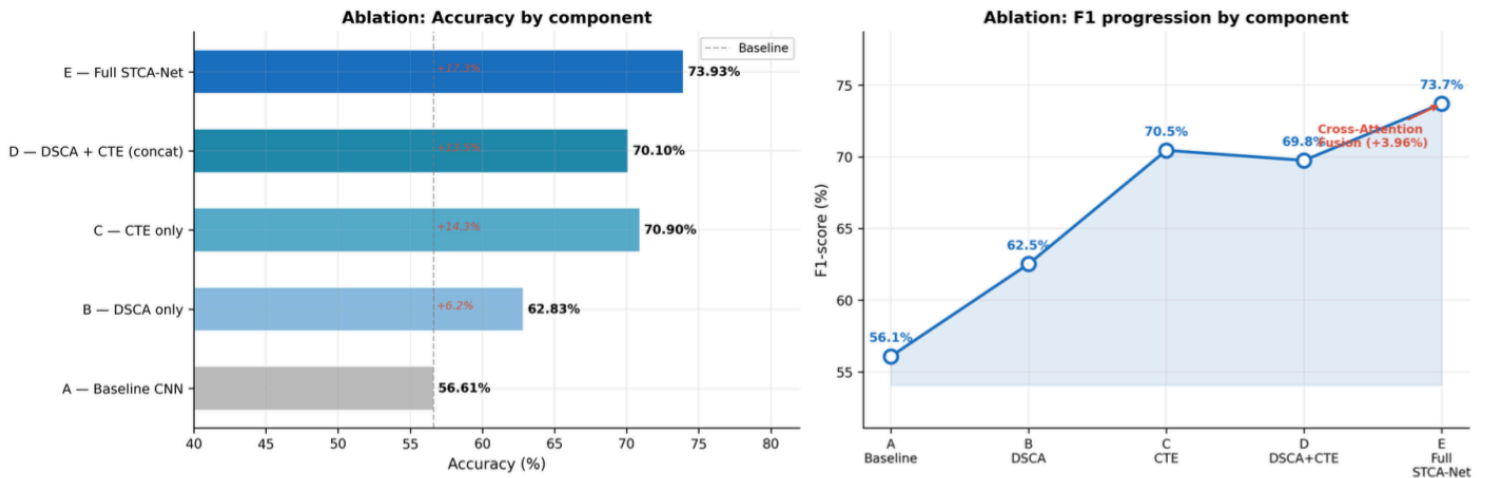


Figure 3

Ablation results on the five-class task (5-fold CV): accuracy by component (left) and F1-score progression (right).