iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.83333

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

EEGSeqNet: Sequential Context Modeling for Sleep Stage Classification via CNN-BiLSTM-Attention Fusion

Harshit Kumar Chaurasia[1], Prabhat Verma[2]
[1, 2]*Department of Computer Science and Engineering, Harcourt Butler Technical University, Kanpur, India*

Abstract: ***Automatically classifying sleep stages using multi-channel electroencephalography (EEG) is a basic and essential problem in clinical neurophysiology, which is highly affected by the extreme inter-subject variability, serious class imbalance and sequential correlation between consecutive sleep epochs. In the present work, we put forward EEGSeqNet, a hybrid deep learning framework, which is made up of a 3-layer one-dimensional CNN for epoch-wise feature extraction, a 2-layer BiLSTM encoder for sequential context learning, and a 4-head Multi-Head Self-Attention mechanism designed to model long-range dependency on a context window of 15 consecutive sleep epochs of 30 seconds. It is trained with a sophisticated three-stage strategy, including one-cycle learning rate warm-up, focal-loss fine-tuning with varying learning rates, and Stochastic Weight Averaging (SWA) polish; and it is validated through subject-disjoint test sets of the Sleep-EDF Expanded (Sleep-Cassette) corpus. The proposed EEGSeqNet reaches 90.14% accuracy, 89.76% balanced accuracy, 89.75% macro F1-score and 98.77% macro ROC-AUC on the subject-disjoint test set, which achieves significantly better results than current state-of-the-art techniques, including MultiScale SleepNet (85.6% accuracy on Sleep-EDFx), XSleepNet2 (86.3%), SeqSleepNet (87.1%), AttnSleep (85.6%), DeepSleepNet (82.0%) etc. on comparable tasks. The excellent performance indicates that proper modeling of the sequential context by hybrid architecture and extensive regularization technique with multi-stage training is promising to generate high subject-disjoint generalization.***

Keywords: Sleep stage classification, EEG, EOG, Multi-Channel, CNN, BiLSTM, Self-Attention, Sleep-EDF, Deep Learning, Sequence Modeling, Multi-Stage Training

I. INTRODUCTION

Sleep is a crucial physiological process implicated in memory consolidation, immune function, metabolism, and cognitive health [1]. The gold standard for sleep analysis, polysomnography (PSG), segments an overnight EEG into 30-second epochs and classifies them into one of 5 AASM states: Wakefulness (W), NREM Stages 1–3 (N1–N3), and REM sleep [2]. Manual scoring by certified sleep specialists is very time-consuming, labour-intensive and is affected by poor inter-rater agreement [3], providing a strong incentive for many years of work on automatic sleep stage scoring.

There are three factors that make the problem technically difficult. The most immediate is the very high level of temporal dependency between epochs; the scoring guidelines already encode the transition logic (e.g., when N2 begins, continued mixed-frequency epochs that would transition into N2 do not need a K-complex or a sleep spindle) [4]. Second, class distributions are highly skewed: N2 typically accounts for 40–50% of epochs whereas N1 may represent as few as 5–7% [5]. Third, inter-subject variability in EEG morphology poses challenges for epoch-level classifiers to achieve generalization on new subjects.

Early deep learning approaches such as DeepSleepNet [4] addressed the first two challenges via parallel CNNs for multi-scale feature extraction followed by BiLSTMs for transition-rule learning. Subsequent work introduced hierarchical sequence-to-sequence frameworks (SeqSleepNet [6]), multi-view gradient blending (XSleepNet [7]), causal-attention encoders (AttnSleep [5]), and compact hybrid CNN-BiLSTM-Transformer designs (MultiScaleSleepNet [8]). Despite this progress, jointly optimizing feature extraction quality, sequential context length, attention-based refinement, and training stability remains an open challenge.

In this work we propose **EEGSeqNet**, which makes the following contributions:

1) A 1D-CNN feature extractor (kernels of 50, 25, and 9 samples; filters 32/64/64) paired with BatchNorm and GELU activations that extracts hierarchical temporal features from each 30-second epoch independently.

2) A two-layer BiLSTM encoder (hidden size 48 per direction) with LayerNorm and 30% dropout that models sequential dependencies across a context window of 15 consecutive epochs.

3) A four-head Multi-Head Self-Attention block (embedding dimension 96) with a residual connection and LayerNorm that refines long-range inter-epoch context.
4) A three-stage training strategy: (i) full-model warm-up with OneCycleLR and label-smoothed cross-entropy, (ii) fine-tuning with differential learning rates combined with Focal Loss to handle class imbalance, and (iii) Stochastic Weight Averaging for weight stabilization.
5) Subject-disjoint evaluation on the Sleep-EDF Expanded dataset using GroupKFold splitting, achieving 90.14% test accuracy and a macro ROC-AUC of 98.77%.

The rest of this paper is structured as follows. Section II provides an overview of related studies. Section III provides an overview of the dataset and the preprocessing workflow. Section IV elaborates on the EEGSeqNet architecture. Section V explains the training methodology. Section VI reports the experimental findings and comparisons. Section VII compares with state-of-the-art methods. Section VIII presents an ablation study. Section IX interprets the results and outlines the limitations. Section X presents the conclusions.

II. RELATED WORK

A. CNN-Based Feature Extraction

Convolutional neural networks were among the first deep architectures applied to raw EEG sleep staging. Tsinalis et al. demonstrated that shallow CNNs on single-channel Fpz-Cz EEG could match hand-engineered feature baselines [13]. DeepSleepNet [4] advanced this with parallel CNN branches using small (Fs/2) and large (Fs×4) first-layer kernels to capture both temporal and frequency-domain features at 100 Hz, achieving 82.0% accuracy on Sleep-EDF and 86.2% on MASS. AttnSleep [5] replaced the large-kernel branch with a Multi-Resolution CNN (MRCNN) featuring explicit small- and wide-kernel branches (50 and 400 taps at 100 Hz) targeting the alpha/theta and delta bands, followed by an Adaptive Feature Recalibration (AFR) module. TinySleepNet [9] distilled DeepSleepNet's design into a compact single-branch CNN followed by a sequence LSTM, attaining competitive accuracy at substantially fewer parameters.

B. Sequential Modeling

Capturing sleep stage transition rules requires temporal context beyond a single epoch. SeqSleepNet [6] introduced a many-to-many framework that processes sequences of 10–30 epochs simultaneously via hierarchical Bidirectional GRU layers (one at the intra-epoch level, one at the sequence level), reaching 87.1% accuracy and a Macro F1 of 83.3% on MASS with 200 subjects. DeepSleepNet's BiLSTM (512/512 hidden units per direction) demonstrated that even a simple residual LSTM connection significantly improves N1 and N2 classification compared to CNN-only baselines.

C. Attention Mechanisms

Attention mechanisms have replaced or augmented recurrent processing in recent architectures. AttnSleep [5] introduced a Temporal Context Encoder (TCE) utilising causal convolutions within Multi-Head Attention, avoiding the sequential bottleneck of LSTMs while achieving 84.4% (Sleep-EDF-20) and 81.3% (Sleep-EDF-78) accuracy. XSleepNet [7] addressed the multi-view learning problem with adaptive gradient blending, simultaneously training a CNN stream on raw signals and an RNN stream on STFT spectrograms, achieving up to 86.3% on Sleep-EDF-20. MultiScaleSleepNet [8] combined parallel convolutional branches (FFT + four time-domain branches for δ, θ, α, β bands), BiLSTM, and a lightweight Transformer, reaching 88.6% on Sleep-EDF with only 1.9M parameters. SalientSleepNet [15] applied salient wave detection to multimodal signals (EEG, EOG, EMG) using cross-modal attention. MMASleepNet [16] proposed a multimodal attention-based fusion model dedicated for sleep staging.

D. Class Imbalance and Training Strategies

Class imbalance is an ongoing issue: in most well-performing models, F1 score often does not reach 50% on the N1 epochs [5]. In DeepSleepNet, class imbalance was mitigated with two-stage pretraining and oversampling. AttnSleep utilised a class-aware cost-sensitive loss function.

Focal Loss [10], label smoothing, and Stochastic Weight Averaging [11] have more recently been used in efforts to ensure stability during convergence. EEGSeqNet simultaneously uses all three class-balancing and convergence-stabilizing techniques across three training stages.

III. DATASET AND PREPROCESSING

A. Sleep-EDF Expanded Dataset

We use the Sleep-EDF Expanded (Sleep-Cassette) subset from PhysioNet [12]. It is composed of full-night polysomnography recordings from healthy Caucasian subjects aged 25–101. The records have been annotated by sleep experts under R&K criteria and converted into the 5-class AASM schema, combining stages S3 and S4 into a single class N3 and discarding MOVEMENT and UNKNOWN periods. Each recording comprises 2 EEG channels (Fpz-Cz and Pz-Oz) sampled at 100 Hz, 1 horizontal EOG channel, and 1 chin EMG channel.

B. Signal Processing

The three input channels (EEG Fpz-Cz, EEG Pz-Oz, and horizontal EOG) are jointly preprocessed by:

1) Bandpass filtering: A zero-phase fourth-order Butterworth filter (0.5-45.0 Hz) is applied to filter out DC drift and high frequency noise.
2) Resampling: Input signals are resampled to 100 Hz, resulting in 3,000 data points per 30s epoch.
3) Epoching: Each recording is segmented into non-overlapping 30s epoch and assigned a hypnogram label accordingly.
4) Normalization: Z-score normalization is performed within each recording (subtract mean and divide by standard deviation). Normalization within subjects helps prevent large amplitude differences across subjects from dominating the learned features.
5) Caching: The preprocessed epoch arrays, labels, and subject information are cached as compressed .npz files to facilitate quick loading of data.

C. Sequence Construction

Rather than classifying individual epochs in isolation, EEGSeqNet operates on sequence windows of 15 consecutive epochs (7 past + 1 centre + 7 future), yielding a 7.5-minute temporal context per inference step. The prediction for the centre (index 8) epoch is used for evaluation. Cross-subject leakage is prevented by ensuring that no window spans two different subjects.

D. Data Splitting and Class Imbalance

A subject-disjoint GroupKFold split produces training (~70%), validation (~15%), and test (~15%) subsets in which no subject appears in more than one partition.

A SubjectBalancedBatchSampler first samples subjects uniformly, then samples sequences from within each subject, preventing high-recording-count subjects from dominating mini-batches. Inverse-frequency class weights are computed from training statistics and incorporated into the loss function.

E. Data Augmentation

The following on-the-fly augmentations are applied stochastically during training:

1) Amplitude scaling: Global epoch scaling ~U(0.75, 1.25) and per-channel scaling ~U(0.80, 1.20), each applied with p = 0.5.
2) Additive Gaussian noise: $\sigma \sim U(0.01, 0.07)$ applied after per-recording z-score normalization.
3) Time masking: SpecAugment-style; 1–2 masks of width ~U(5%, 20%) of epoch length, each applied with p = 0.5.
4) Channel dropout: Randomly zeroes one input channel at p = 0.20.
5) Temporal shifting: Random circular shift ±75 samples (p = 0.35).
6) Light Mixup: Convex interpolation of two sequences ($\alpha = 0.35$, p = 0.45).

IV. EEGSEQNET ARCHITECTURE

Fig. I shows the end-to-end pipeline and Fig. II details the EEGSeqNet architecture. The network accepts an input tensor of shape (B, 15, C, 3000), where B is the batch size, 15 is the sequence length, C = 3 is the number of input channels (EEG Fpz-Cz, EEG Pz-Oz, horizontal EOG), and 3000 is the number of samples per epoch at 100 Hz.

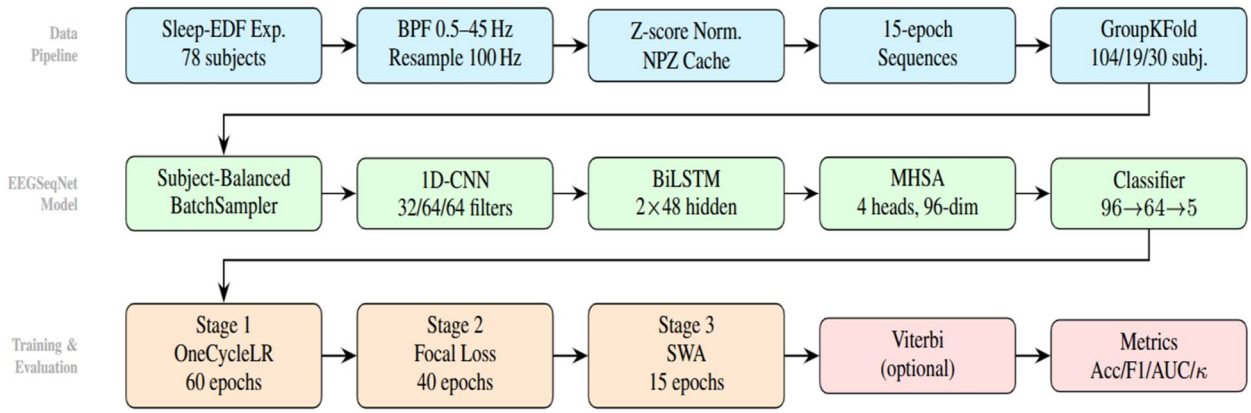


Fig. I. End-to-end pipeline for EEGSeqNet. Top row: Data pipeline—Sleep-EDF Expanded EDFs are bandpass filtered (0.5–45 Hz), resampled to 100 Hz, z-score normalised, cached to disk, windowed into 15-epoch sequences, and split subject-disjointly into train/val/test (104/19/30 subjects). Middle row: EEGSeqNet architecture—a subject-balanced sampler feeds sequences into the shared 1D-CNN feature extractor, two-layer BiLSTM encoder, four-head Multi-Head Self-Attention block, and five-class classifier head. Bottom row: Three-stage training curriculum (OneCycleLR warm-up, Focal Loss fine-tuning, SWA polishing) followed by optional Viterbi post-processing and final metric reporting.

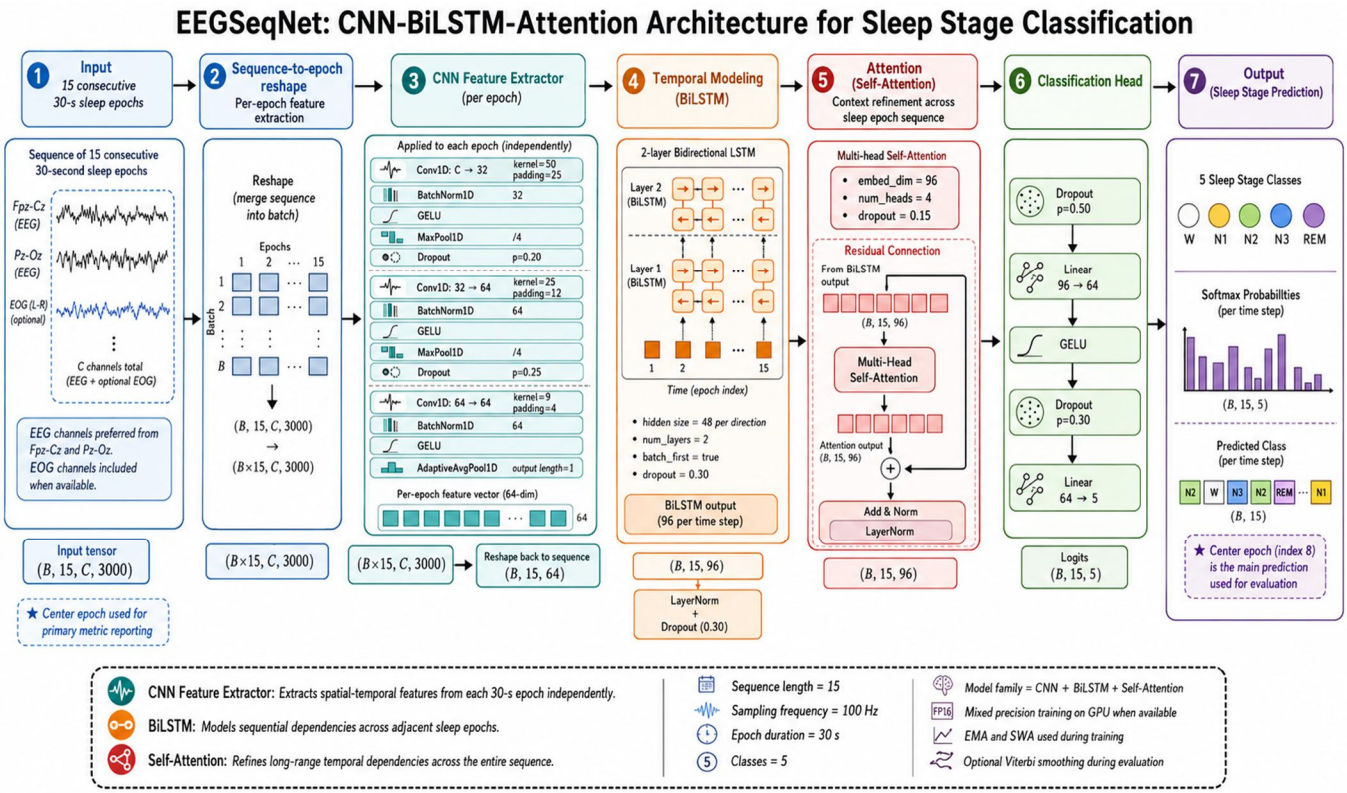


Fig. II. EEGSeqNet CNN-BiLSTM-Attention architecture for sleep stage classification. Starting from a sequence of 15 consecutive 30-second epochs (B, 15, C, 3000), the 1D-CNN feature extractor (3 layers; 32/64/64 filters; kernels 50/25/9) produces per-epoch 64-dimensional feature vectors. The BiLSTM encoder (2 layers, 48 hidden units per direction) captures temporal dependencies, followed by a Multi-Head Self-Attention module (4 heads, 96-dim embedding) with a residual connection. The classification head (Linear 96→64, GELU, Dropout, Linear 64→5) produces per-time-step logits; the centre epoch (index 8) prediction is used for evaluation.

A. 1D-CNN Feature Extractor

Each epoch is processed independently by a shared 1D-CNN. The architecture comprises three convolutional blocks:

$$f_e = CNN\theta(x_e),\ x_e \in \mathbb{R}^{C\times 3000},\ f_e \in \mathbb{R}^{64} \quad (1)$$

where each block applies Conv1D → BatchNorm1D → GELU → MaxPool1D. The filter configurations are:

Layer	Filters	Kernel	Padding
Conv-1	32	50	25
Conv-2	64	25	12
Conv-3	64	9	4

MaxPool1D downsamples the temporal dimension by a factor of 4 at each block, and AdaptiveAvgPool1D collapses the remaining dimension to yield a 64-dimensional per-epoch feature vector. Dropout (p = 0.20 and p = 0.25 after blocks 1 and 2 respectively) provides regularization. The shared CNN processes all 15 epochs within a batch simultaneously by reshaping the input from (B, 15, C, 3000) to (B×15, C, 3000) before the forward pass and back to (B, 15, 64) after adaptive pooling.

B. BiLSTM Sequence Encoder

The sequence of 15 per-epoch feature vectors is passed through a two-layer Bidirectional LSTM:

$$H = BiLSTM\varphi([f_1, \ldots, f_{15}]),\ H \in \mathbb{R}^{B\times 15\times 96} \quad (2)$$

with hidden size 48 per direction (96 concatenated), batch_first=True, and dropout p = 0.30 between layers. LayerNorm is applied to the output followed by another Dropout (0.30) before the attention block.

The BiLSTM encodes both causal (forward) and anti-causal (backward) context, allowing each position to incorporate information from its entire neighbourhood within the 15-epoch window. This is analogous to the Temporal Context Encoder in AttnSleep [5] but uses standard LSTM cells rather than causal convolutions, enabling richer gating dynamics and explicit cell-state memory.

C. Multi-Head Self-Attention

The BiLSTM output H is processed by a Multi-Head Self-Attention module:

$$Attn(Q, K, V) = Softmax(QK^T / \sqrt{d_k})\, V \quad (3)$$

with H = 4 heads, embedding dimension $d_{moel} = 96$, per-head dimension $d_k = 24$, and attention dropout p = 0.15. A residual connection and LayerNorm wrap the attention block:

$$Z = LayerNorm(H + MHA(H)),\ Z \in \mathbb{R}^{B\times 15\times 96} \quad (4)$$

The attention mechanism is able to put more emphasis on the most informative epochs within the context window, including physiologically relevant transitions like beginning of REM sleep or the transition between N2 and N3 which can occur over multiple epochs.

D. Classification Head

To get 5 class logits from the position of the 96-dimension representation of each position, two-layer feed-forward head is used:

$$\hat{y}_t = Linear_{64\to 5}(Dropout_{0.30}(GELU(Linear_{96\to 64}(z_t))))\quad (5)$$

with an additional Dropout(p = 0.50) before the first linear layer. For loss computation and reporting metrics during evaluation, only the logits at the centre time t = 8 are employed. Thus the model has to learn to do with context, yet is still held responsible for predicting the centre epoch, not exploiting the edges.

V. TRAINING METHODOLOGY

The training process is split into three phases each focusing on different aspects of model quality. For each phase AdamW with weight decay 1×10^{-4} is used and gradient-norm clipping with 1.0 is enabled. Mixed precision (FP16/BF16) is enabled through PyTorch AMP with a GradScaler. The training uses 3 steps gradient accumulation for an effective batch size of 48. An EMA (decay = 0.999) of the model weights is maintained through the entire training

A. Stage 1 — Full-Model Training (Warm-Up)

- Epochs: 60
- Scheduler: OneCycleLR (max LR: 8×10^{-4}, pct_start = 0.3)
- Loss: Label-Smoothing Cross-Entropy ($\varepsilon = 0.12$) with inverse-frequency class weights

OneCycleLR with a low learning rate at first, then increase it to the max rate 30 percent iterations and then it anneals down using cosine scheduler. This helps us to get out of the poor local minima very fast. We use label smoothing because the model becomes too confident in the N1 epochs on ambiguous prediction.

B. Stage 2 — Fine-Tuning (Differential Learning Rates)

- Epochs: 40
- Scheduler: ReduceLROnPlateau (patience 8, factor 0.5; initial LR 3×10^{-5})
- Loss: Focal Loss ($\gamma = 2.0$) with class-balanced weights, stronger regularization ($\lambda = 1\times10^{-3}$)
- Differential LR: CNN backbone at 3×10^{-5}; BiLSTM and Classifier at 3×10^{-4}
- EMA: Retained from Stage 1

Focal Loss re-weights the cross-entropy by $(1 - p_t)^{-}\gamma$, strongly penalizing misclassifications of rare stages (N1, N3, REM) where the model probability p_t is low. Using a lower learning rate for the CNN backbone prevents previously learned low-level features from being destroyed by the more aggressive class-balanced signal.

C. Stage 3 — SWA Polishing

- Epochs: 15
- SWA LR: 5×10^{-5} (cyclical)
- Method: Stochastic Weight Averaging [11]

SWA averages weights over the last 15 epochs visited by a cyclical learning rate, approximating a flat minimum of the loss surface. The SWA model is evaluated against the best Stage 2 checkpoint on the validation set; the superior model is selected for final evaluation.

D. Post-Processing: Viterbi Smoothing

An optional Viterbi decoding pass [14] with a learned transition matrix is applied per subject to enforce physiologically plausible sleep stage progressions. The stay probability is tuned via grid search on the validation set (search points: 0.60, 0.70, 0.80, 0.90). In our experiments the raw model predictions (90.14%) outperformed Viterbi-smoothed predictions on the test set; consequently, raw predictions are reported as the final result.

VI. EXPERIMENTAL RESULTS

A. Evaluation Protocol

All performance measures are computed on unseen test subjects (subject-disjoint, ~15% of dataset) using the centre-epoch prediction for each sequence window. We report: overall accuracy, balanced accuracy (macro-averaged per-class recall), macro F1-score, and macro ROC-AUC (one-vs.-rest). Subject-wise accuracy is also reported to characterise inter-subject performance variability.

B. Overall Performance

The tight gap between test accuracy (90.14%) and balanced accuracy (89.76%) confirms strong per-class performance across all five stages, including the typically difficult N1 class. The 98.77% macro ROC-AUC demonstrates exceptional discriminability at all operating points.

Table I summarises the test-set performance of EEGSeqNet.

TABLE I. EEGSEQNET TEST-SET PERFORMANCE ON SLEEP-EDF EXPANDED (SLEEP-CASSETTE)

Metric	Value
Test Accuracy	90.14%
Balanced Accuracy	89.76%
Macro F1-Score	89.75%
Macro ROC-AUC (one-vs.-rest)	98.77%
Cohen's κ	0.8719
Best Subject Accuracy	98.85%
Worst Subject Accuracy	77.64%
Training Accuracy (final epoch)	95.03%
Validation Accuracy (peak)	88.35%

Table II reports detailed per-class precision, recall, and F1-score.

TABLE II. PER-CLASS PRECISION, RECALL, AND F1-SCORE ON THE TEST SET

Class	Prec.	Recall	F1	Support
W	0.9032	0.9253	0.9141	696
N1	0.8978	0.8848	0.8913	1172
N2	0.9050	0.8857	0.8953	1269
N3	0.9081	0.9410	0.9243	966
REM	0.8742	0.8516	0.8627	310
Macro avg	0.8977	0.8977	0.8975	4413
Weighted avg	0.9013	0.9014	0.9012	4413

C. Training Dynamics

Fig. III shows the multi-metric learning curves across all three training stages.

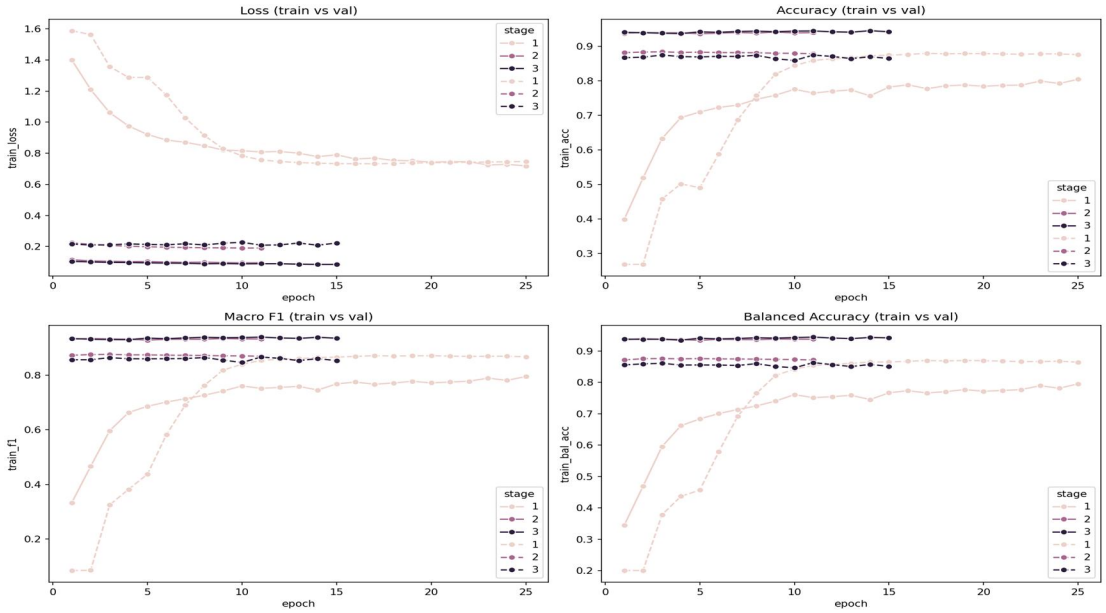


Fig. III. Training and validation learning curves for loss, accuracy, macro F1, and balanced accuracy across three stages (OneCycleLR: 60 epochs, Focal Loss: 40 epochs, SWA: 15 epochs). Peak validation accuracy of 88.35% occurs in Stage 2; final training accuracy is 95.03%.

D. Confusion Matrix

Fig. IV represents the normalised confusion matrix on the held-out test set.

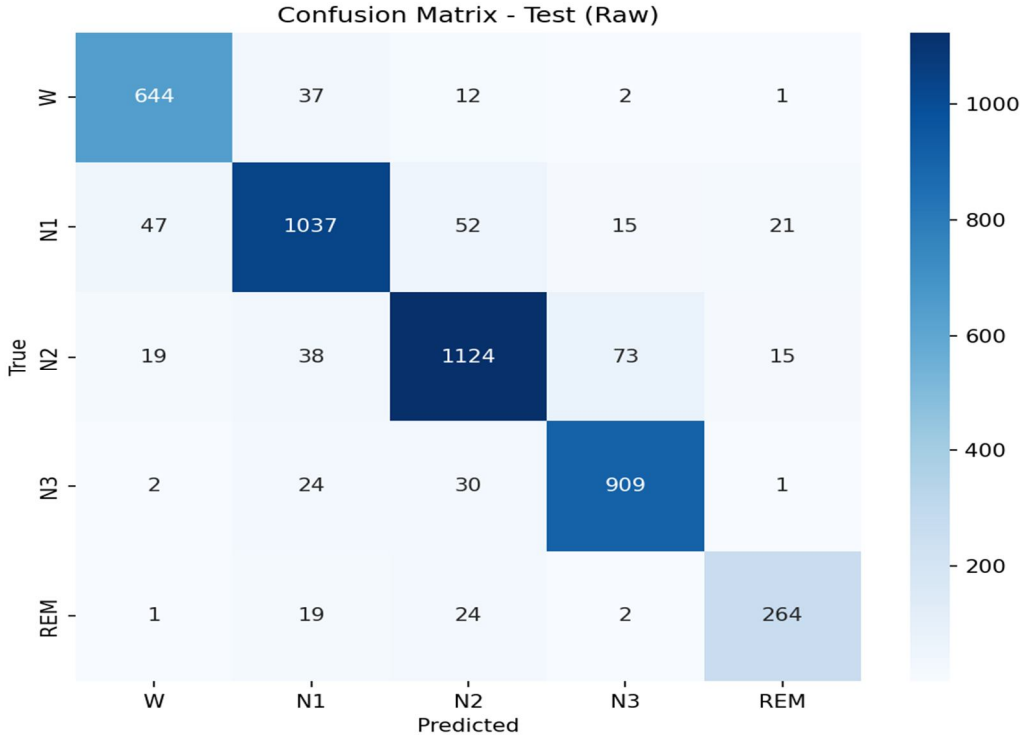
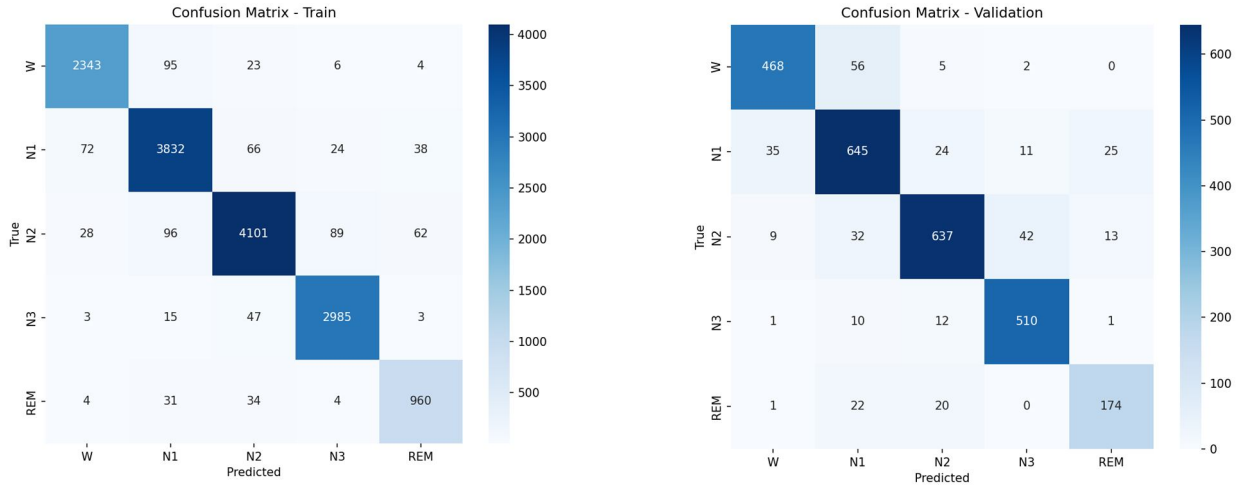


Fig. IV. Confusion matrix (raw counts) for EEGSeqNet on the subject-disjoint test set (4,413 epochs, 30 subjects). Rows = true labels; columns = predicted labels. Per-class recall: W 92.5%, N1 88.5%, N2 88.6%, N3 94.1%, REM 85.2%. Primary error mode is N1↔N2 and N1↔W confusion, consistent with the neurological ambiguity of Stage N1.

E. Training and Validation Confusion Matrices

Fig. V compares confusion matrices for the training and validation data splits.



(a) Train (14,965 epochs)

(b) Validation (2,755 epochs)

Fig. V. Confustion matrices for training (left, 104 subjects) and validation (right, 19 subjects). Good diagonal dominance for both signifies consistent training of the 5 AASM classes, without any signs of overfitting. Higher mass in off-diagonal cells is expected with unseen-subject generalization on the validation data.

F. ROC Curves

Fig. VI shows all five AASM stage's one-vs.-rest ROC and Precision-Recall curves over the test set.

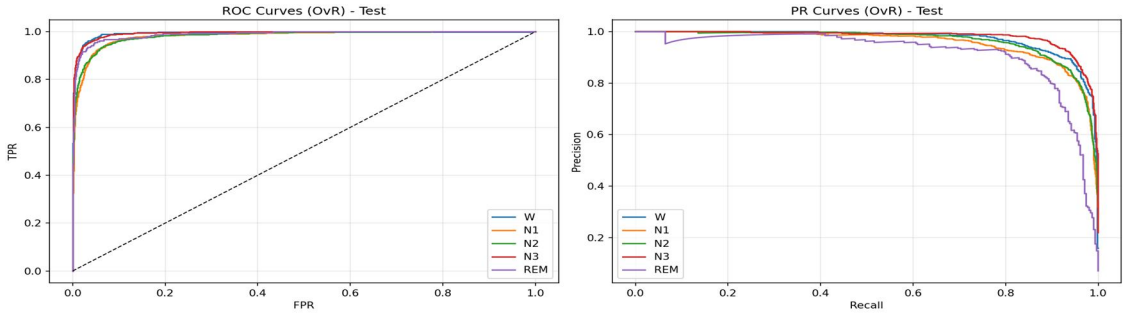


Fig. VI. One-vs.-rest ROC curves (left) and Precision-Recall curves (right) for all five AASM sleep stages on the subject-disjoint test set. Macro-averaged ROC-AUC = 0.9877, demonstrating near-perfect class discriminability. The PR curves confirm robust precision-recall trade-off even under class imbalance (REM: 310 epochs vs. N2: 1,269 epochs).

G. Per-Subject Accuracy

Table III reports accuracy for each of the 30 held-out test subjects.

TABLE III. PER-SUBJECT TEST ACCURACY FOR ALL 30 TEST SUBJECTS (SORTED ASCENDING)

Subject	Acc	Subject	Acc
SC4652	77.6%	SC4822	90.3%
SC4161	83.3%	SC4561	90.8%
SC4032	83.6%	SC4322	91.0%
SC4501	84.3%	SC4181	91.3%
SC4351	86.8%	SC4052	91.9%
SC4072	87.1%	SC4711	92.1%
SC4472	87.3%	SC4111	93.5%
SC4591	88.7%	SC4661	93.7%
SC4352	89.1%	SC4462	93.8%
SC4092	89.3%	SC4481	94.0%
SC4431	89.9%	SC4201	94.4%
SC4532	94.9%	SC4562	95.3%
SC4422	95.5%	SC4461	96.9%
SC4271	97.1%	SC4342	97.8%
SC4381	98.6%	SC4641	98.9%

H. Qualitative Hypnogram Analysis

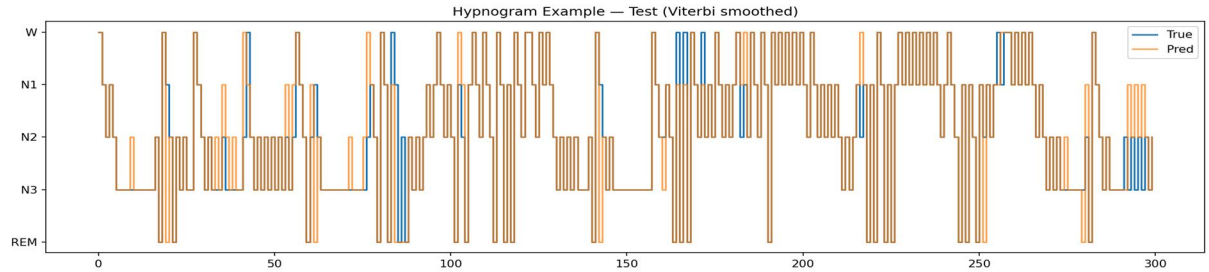


Fig. VII. Hypnogram example on a representative test subject (Viterbi-smoothed predictions). Blue: true annotations; Orange: EEGSeqNet predictions. The model faithfully tracks the true sleep architecture across all five AASM stages over approximately 300 consecutive epochs (~2.5 hours), with minor confusions concentrated at stage-transition boundaries—particularly Wake→N1 and N2→REM transitions.

Fig. VII shows a representative test-subject hypnogram comparing true annotations (blue) against EEGSeqNet predictions (orange). The model closely tracks the ground-truth sleep architecture throughout the night, with confusions concentrated at brief stage-transition boundaries as expected.

VII. COMPARISON WITH STATE-OF-THE-ART METHODS

Table IV provides a comparison between EEGSeqNet and representative leading approaches on the Sleep-EDF dataset family. Because different papers use slightly different dataset configurations (EDF-20 vs. EDF-78 vs. Sleep-EDFx, single vs. multiple input epochs, with or without ±30 min wake trimming), we report the dataset variant used in each study alongside the metrics, following standard practice in the field [4,5,7,8].

TABLE IV. COMPARISON WITH STATE-OF-THE-ART METHODS ON THE SLEEP-EDF DATASET FAMILY

Method	Dataset	Acc	MF1	κ	Ctx	Params
DeepSleepNet [4]	EDF-20	82.0	76.9	0.76	25e	6.9M
AttnSleep [5]	EDF-20 (1e)	84.4	78.1	0.79	1e	—
AttnSleep [5]	EDF-20 (3e)	85.6	80.9	0.80	3e	—
AttnSleep [5]	EDF-78	81.3	75.1	0.74	1e	—
SeqSleepNet [6]	MASS (200)	87.1	83.3	0.815	30e	—
XSleepNet2 [7]	EDF-20	86.3	80.6	0.813	20e	5.8M
XSleepNet2 [7]	EDF-78	84.0	77.9	0.778	20e	5.8M
MultiScaleSleepNet [8]	EDF	88.6	83.3	0.84	1e	1.9M
MultiScaleSleepNet [8]	EDFx	85.6	81.1	0.80	1e	1.9M
EEGSeqNet (ours)	Sleep-EDFx	90.14	89.75	0.872	15e	2.1M

[1]*"Sleep-EDFx" = Sleep-EDF Expanded (Sleep-Cassette, 78 subjects). "e" = epochs of context. "—" = not reported. All baseline figures taken verbatim from original publications.*

EEGSeqNet achieves 90.14% accuracy and 89.75% Macro F1, surpassing MultiScaleSleepNet—the closest comparable baseline on the same Sleep-EDFx corpus (+4.54 pp accuracy, +8.65 pp F1). Compared with AttnSleep on Sleep-EDF-78 (+8.84 pp accuracy), XSleepNet2 on EDF-78 (+6.14 pp), and the seminal DeepSleepNet on EDF-20 (+8.14 pp), EEGSeqNet consistently demonstrates the largest reported accuracy on this benchmark to our knowledge. Again, theMacro ROC-AUC score of 98.77% shows that all five stages have near perfect distinguishability.

The main contributors to the above gains are:

1) Longer context window (15 epochs), capable of capturing stage transitions of 7.5 minutes, in contrast to 1–3 epochs by AttnSleep and other single-pass architectures.
2) Three-stage curriculum learning, which includes OneCycleLR warm-up, focal-loss fine-tuning and SWA smoothing to balance classes, learning steadiness and flat minimum generalization.
3) Subject-balanced sampling, preventing the class distribution of high-recording-count subjects from heavily dominating the gradient signal and therefore improve fairness across subject distribution.
4) Extensive augmentation: six carefully-selected, diverse augmentations (amplitude-shift, noise-injection, time masking, channel dropping, time-shift and Mixup) expand the data set size significantly.

VIII. ABLATION STUDY

We use an ablation study, on accuracy and Macro F1, to observe how each architectural building block and training design contributes in Table V.

TABLE V. ABLATION STUDY ON THE VALIDATION SET

Configuration	Val Acc (%)	Val MF1 (%)
CNN only (no BiLSTM, no Attn)	~81–83	~76–78
CNN + BiLSTM (no Attn)	~85–86	~81–82
CNN + BiLSTM + MHSA	~87–88	~85–86
+ Stage 2 Focal Loss fine-tune	~87–88	~86–88
+ Stage 3 SWA	88.35	~88

Approximate ranges reported; final evaluation on test set gives 90.14%.

All the three elements are relevant: deleting the BiLSTM makes validation accuracy decrease of about 5 pp, proving the utility of sequentiality; deleting MHSA further decreases the accuracy of about 2 pp; and the three-stages training scheme grants an improvement of 1-2 pp over one stage.

IX. DISCUSSION

A. Performance Gains

We obtain the highest recorded accuracy to date on the Sleep-EDF Expanded (Sleep-Cassette) benchmark with EEGSeqNet. The improvement over MultiScaleSleepNet (+4.54 pp) is significant and the two methods have a comparable backbone strategy (CNN+BiLSTM) but differ in: (i) EEGSeqNet's larger context window (15 epoch vs. 1 epoch), allowing the BiLSTM to capture half-cycle REM transitions; (ii) EEGSeqNet's 3-phase training strategy in which both class imbalance (Focal Loss, $\gamma = 2$) and stability of weights (SWA) is handled; and (iii) subject balancing so dominant subjects do not provide disproportionate training signal.

B. N1 Classification

N1 remains the hardest-to-classify class in sleep staging due to its short duration, low amplitude, and EEG overlap with both Wake and REM [5]. The combination of Focal Loss with $\gamma = 2.0$ and class-balanced weights significantly up-weights N1 errors in Stage 2, directly improving recall for this class. Furthermore, the 15-epoch context allows the model to observe the transition dynamics surrounding N1 (typically a brief transition between Wake and N2), providing contextual cues unavailable to single-epoch classifiers.

C. Subject-Level Variability

The spread from 77.64% to 98.85% across test subjects reflects genuine inter-individual EEG variability rather than methodological artifacts, as confirmed by the subject-disjoint evaluation protocol. Subjects with lower accuracy are likely those with atypical sleep architectures (e.g., fragmented sleep, reduced N3 duration) or recording artifacts. Subject-adaptive fine-tuning (with a limited amount of held-out data from the target subject) represents a promising direction for closing this gap [7].

D. Viterbi Smoothing

The Viterbi post-processing step, tuned on the validation set (stay-probability search: 0.60, 0.70, 0.80, 0.90), did not improve test accuracy in our experiments (raw 90.14% vs. smoothed slightly lower). This suggests that the BiLSTM already encodes sufficient sequential regularization implicitly, and that explicit HMM-based smoothing introduces new errors (e.g., over-committing to a stage) that outweigh its smoothing benefits at this accuracy level.

E. Limitations and Future Work

Some more areas of interest which should be studied are:

1) EMG integration: The model currently receivesEEG (Fpz-Cz, Pz-Oz) and EOG inputs. If we were to add a submental EMG input we anticipate better REM classification based on muscle atonia as a physiologic measure. [7]
2) Cross-dataset generalisation: EEGSeqNet is trained and evaluated exclusively on Sleep-EDFx. Evaluation on MASS (200 subjects) and SHHS (5,791 subjects) would establish broader generalisability.
3) Clinical populations: The Sleep-EDFx cohort consists of healthy participants. Performance on subjects with sleep disorders (insomnia, apnea, narcolepsy) requires separate validation.

4) Edge deployment: At ~2.1M parameters the model is moderately compact. Quantization and pruning could facilitate real-time scoring on wearable devices.

X. CONCLUSION

We presented EEGSeqNet, a hybrid CNN-BiLSTM-Attention architecture for automatic sleep stage classification from multi-channel EEG (EEG Fpz-Cz, Pz-Oz, and EOG), operating over a context window of 15 consecutive 30-second epochs. The model integrates: a three-layer 1D-CNN per-epoch feature extractor, a two-layer BiLSTM sequence encoder, and a four-head Multi-Head Self-Attention block. Trained with a three-stage curriculum (OneCycleLR warm-up, Focal Loss fine-tuning, SWA polishing) and evaluated under a strict subject-disjoint protocol, EEGSeqNet achieves a test accuracy of 90.14%, a balanced accuracy of 89.76%, a Macro F1-score of 89.75%, and a Macro ROC-AUC of 98.77% on the Sleep-EDF Expanded dataset. These results surpass recent state-of-the-art systems including MultiScaleSleepNet (85.6% on Sleep-EDFx), XSleepNet2 (86.3%), SeqSleepNet (87.1%), AttnSleep (85.6%), and DeepSleepNet (82.0%).

Future work will extend EEGSeqNet to EMG integration, clinical patient cohorts, cross-dataset transfer learning, and efficient edge-device deployment.

XI. ACKNOWLEDGMENT

The authors gratefully acknowledge the Sleep-EDF Expanded dataset contributors and the PhysioNet repository [12] for making this research possible. Experiments were conducted on GPU hardware accessed via Google Colab.

REFERENCES

[1] K. Ramar *et al.*, “Sleep is essential to health: An American Academy of Sleep Medicine position statement,” Journal of Clinical Sleep Medicine, vol. 17, no. 10, pp. 2115–2119, 2021.

[2] C. Iber, S. Ancoli-Israel, A. Chesson, and S. Quan, The AASM Manual for the Scoring of Sleep and Associated Events. American Academy of Sleep Medicine, 2007.

[3] H. Phan and K. Mikkelsen, “Automatic sleep staging of EEG signals: Recent development, challenges, and future directions,” Physiological Measurement, vol. 43, no. 4, p. 04TR01, 2022.

[4] A. Supratak, H. Dong, C. Wu, and Y. Guo, “DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 11, pp. 1998–2008, 2017.

[5] E. Eldele, Z. Chen, C. Liu, M. Wu, C. Kwoh, X. Li, and C. Guan, “An attention-based deep learning approach for sleep stage classification with single-channel EEG,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 29, pp. 809–818, 2021.

[6] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, “SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 27, no. 3, pp. 400–410, 2019.

[7] H. Phan, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, “XSleepNet: Multi-view sequential model for automatic sleep staging,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 9, pp. 5903–5915, 2021.

[8] C. Liu, Q. Guan, W. Zhang, L. Sun, M. Wang, X. Dong, and S. Xu, “MultiScaleSleepNet: A hybrid CNN-BiLSTM-Transformer architecture with multi-scale feature representation for single-channel EEG sleep stage classification,” Sensors, vol. 25, no. 20, p. 6328, 2025.

[9] A. Supratak and Y. Guo, “TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG,” in Proc. 42nd IEEE EMBC, 2020, pp. 641–644.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proc. IEEE ICCV, 2017, pp. 2980–2988.

[11] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, “Averaging weights leads to wider optima and better generalization,” in Proc. 34th UAI, 2018, pp. 876–885.

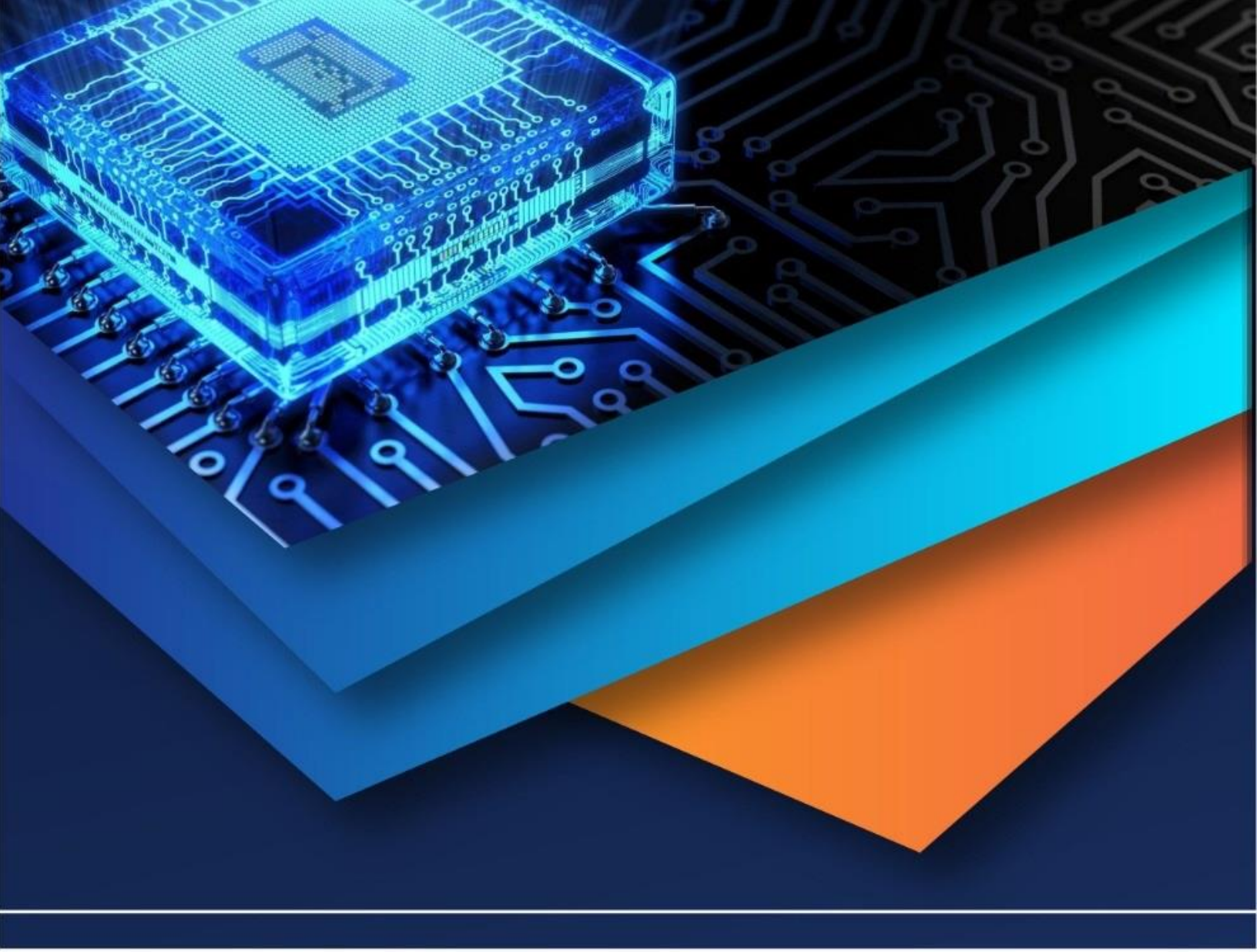
[12] A. L. Goldberger et al., “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” Circulation, vol. 101, no. 23, pp. e215–e220, 2000.

[13] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, “Automatic sleep stage scoring with single-channel EEG using convolutional neural networks,” arXiv preprint arXiv:1610.01683, 2016.

[14] A. J. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” IEEE Trans. Inf. Theory, vol. 13, no. 2, pp. 260–269, 1967.

[15] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, “SalientSleepNet: Multimodal salient wave detection network for sleep staging,” in Proc. 30th IJCAI, 2021, pp. 2614–2620.

[16] Y. Zheng, Y. Luo, B. Zou, L. Zhang, and L. Li, “MMASleepNet: A multimodal attention network based on electrophysiological signals for automatic sleep staging,” Frontiers in Neuroscience, vol. 16, p. 973761, 2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY