Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 10 April 2026
Revised 20 May 2026
Accepted 15 June 2026

Natural Resources for Human Health 2026, Vol. 6 (3s): 945–956

KEYWORDS:
sleep stage classification, retrieval augmented generation, explainable artificial intelligence, electroencephalography, knowledge grounded inference, deep learning

RAGSleepNet: Dual-Store Retrieval for Evidence-Grounded Sleep Staging from Single-Channel EEG

Deepika[1], Rajesh I S[2]*, Kiran Babu T S[3], Chandrappa S[4], Ashwini R[5], A. Shameem Fathima[6]

[1]PES College of Engineering, Mandya, Karnataka, India. Email: bidrideepika1982@gmail.com
[2]Department of AI&ML, BMS Institute of Technology & Management, Bengaluru, Karnataka, India. Email: rajeshaiml@bmsit.in
[3]Associate Professor, Dept of CSE, Gopalan College of Engineering, India. Email: kirangvit@gmail.com
[4]Department of Computer Science and Engineering (Data Science), Nagarjuna College of Engineering and Technology, Bengaluru, Karnataka, India. Email: dr.chandru@ncetmail.com
[5]Department of Computer Science and Engineering, East West Institute of Technology, Bangalore, Karnataka, India. Email: ashwini.ramegowda@gmail.com
[6]Associate Professor II, Dept. OF CSE, R.M.K. College OF Engineering and Technology, Tamil Nadu, India. ORCID: https://orcid.org/0000-0002-1460-8231. Email: shameemfathima81@gmail.com
*Corresponding Author: Rajesh I S (rajeshaiml@bmsit.in)

ABSTRACT: Automatic sleep stage classification now approaches expert agreement on single channel electroencephalography, yet the resulting decisions remain opaque and cannot be audited against the rules clinicians apply. This work introduces RAGSleepNet, a retrieval augmented and explainable staging framework in which every decision is produced together with the evidence supporting it. A dual scale convolutional encoder followed by a transformer context encoder maps each 30 s epoch, with its twenty surrounding epochs, into a contextual query. The query addresses two retrieval branches: a non parametric exemplar memory holding embeddings of all scored training epochs, and a store of 12 clinical knowledge cards whose predicates are soft functions of 17 interpretable descriptors derived from American Academy of Sleep Medicine criteria. A gating network estimates how far the parametric head, the retrieved exemplars and the cards should be trusted for each epoch, and the three posteriors are combined in a log linear mixture. Subject independent five fold cross validation on 39 nights from 20 subjects of the Sleep-EDF Expanded database, comprising 42,307 scored epochs, yields 82.44 percent accuracy, 75.14 percent macro F1 and a Cohen kappa of 0.7588. Band stop probing shows that removing the frequency support cited by the card lowers the posterior of the predicted stage by 0.271, against 0.066 for a random band, confirming that the explanations are faithful.

1. INTRODUCTION

Polysomnography is the reference examination for sleep disorders, and its interpretation begins by assigning a sleep stage to every 30 s epoch. Manual scoring under the American Academy of Sleep Medicine criteria is slow and shows inter scorer agreement that rarely exceeds a Cohen kappa of 0.80, so its automation has been studied intensively [1], and models on a single electroencephalography derivation now approach human agreement using convolutional, recurrent, graph and self attention architectures [2], [3], [4], [5].

Clinical adoption is limited by something accuracy does not capture. A technologist who disagrees with an automatic label needs to know why it was produced, in the vocabulary of the scoring manual: a spindle train, the fraction of the epoch occupied by slow wave activity, the attenuation of the posterior dominant rhythm. Attribution maps and attention weights describe where the network looked rather than which criterion was satisfied, and their faithfulness is frequently unverified [6], [7].

Retrieval augmented generation offers a mechanism for this. In language modelling a retrieval component supplies documents the generator conditions on, so the output can be traced to identifiable sources, and the principle transfers to clinical question answering with gains in accuracy and verifiability [8], [9]. Applying it to physiological time series requires two decisions with no counterpart in text: what constitutes a retrievable document for a 30 s epoch, and how retrieved evidence should combine with a parametric classifier that is usually stronger than any single retrieved item.

This paper answers both. The retrievable corpus is split into two stores: an exemplar memory holding the contextual embedding and expert label of every training epoch, so a decision can be justified by pointing at previously scored epochs that resemble the current one, and a store of clinical cards, each encoding one criterion as soft predicates over interpretable descriptors, so a decision can be justified by naming the criterion it satisfies. A gating network, estimated on nights disjoint from the training and test partitions, weights each branch, and the three posteriors are merged in a log linear mixture. The system is denoted RAGSleepNet and is illustrated in Figure 1.

Fig. 1 Complete RAGSleepNet pipeline with measured outputs

The contributions are threefold. A dual store retrieval mechanism is formulated for sleep staging in which exemplar evidence and codified clinical knowledge are retrieved with one contextual query and treated as contributors to the decision rather than as post hoc explanations. A confidence conditioned gate discounts the retrieval branches where the memory is unreliable, letting a non parametric component contribute alongside a strong backbone instead of overriding it. The explanations are then validated quantitatively by band stop probing, under a subject independent protocol in which memory and gate are built only from training subjects.

2. RELATED WORK

Automatic sleep staging on single channel electroencephalography was established as a deep learning problem by models combining a convolutional representation learner with a recurrent sequence model. Attention based designs augment recurrence with self attention over neighbouring epochs and improve recognition of stage N1 [2], multi view models process the raw waveform and its time frequency image in parallel [3], fully convolutional networks trained across cohorts generalise to unseen setups [4], and long sequence formulations extend the receptive field to the whole night [10].

Graph formulations learn adjacency structures over electrodes or epochs that transfer across cohorts [5], multimodal variants fuse electroencephalography with electrooculography and electromyography through attention or salient wave detection [11], [12], [13], and fusion pipelines combine handcrafted descriptors with learned representations [14]. Transfer learning across heterogeneous montages allows one scorer to be reused at scale [15], sub epoch context networks exploit structure below the 30 s grid [16], and lightweight designs reduce the parameter count and add uncertainty estimates that flag epochs for manual review [17]. Evaluation protocol is itself a subject of study, since mixed cohort settings expose generalisation failures that single cohort splits hide [18], [19].

Interpretability has followed two directions: attaching attribution methods such as layer wise relevance propagation to a trained network [7], or building inspectable attention and calibrated uncertainty into the architecture [20]. Both explain a decision in terms of the model rather than the scoring manual, and neither cites previously scored epochs. Following the argument that high stakes decisions should rest on inspectable reasoning rather than post hoc rationalisation [6], the position taken here is that evidence should be retrieved and should participate in the decision [8], [9]. To the best of our knowledge this is the first dual store retrieval formulation, combining scored exemplars and codified criteria, for sleep stage classification.

3. PROPOSED METHODOLOGY

3.1 Problem formulation and preprocessing

Let a recording be a sequence of N epochs, each a vector $x_t \in \mathbb{R}^T$ from the Fpz-Cz derivation sampled at 100 Hz over 30 s, so $T = 3000$, with expert label $y_t \in \mathcal{C} = \{W, N1, N2, N3, REM\}$. Staging estimates the posterior $p(y_t \mid x_{t-L/2:t+L/2})$ over a context window of L epochs. Amplitude differences between nights are removed by a robust per recording transformation

$$\tilde{x}_t = \mathrm{clip}\left(\frac{x_t - \mathrm{med}(X)}{\mathrm{IQR}(X)+\epsilon}, -20, 20\right) \tag{1}$$

where $\mathrm{med}(X)$ and $\mathrm{IQR}(X)$ are the median and interquartile range of all samples of the recording. The transformation uses no label information and is applied identically to training and test nights.

3.2 Dual scale epoch encoder

Sleep microstructure spans two time scales: K complexes and spindles occupy fractions of a second, while the rhythms defining the deeper stages need several seconds. One receptive field cannot serve both, so the encoder uses a fine branch with a first kernel of 50 samples at stride 6 and a coarse branch with a first kernel of 400 samples at stride 50,

$$h_t^f = \mathcal{F}_{\theta_f}(\tilde{x}_t), \qquad h_t^c = \mathcal{F}_{\theta_c}(\tilde{x}_t) \tag{2}$$

where each branch stacks convolution, batch normalisation, GELU activation and max pooling. The responses are pooled to a common length, concatenated and projected to a d dimensional embedding,

$$e_t = \mathrm{LN}\left(W_p\,[\,\mathrm{pool}(h_t^f) \parallel \mathrm{pool}(h_t^c)\,] + b_p\right) \tag{3}$$

with $d = 128$, where LN is layer normalisation and $\parallel$ is concatenation. The encoder is shared by all epochs of the context window.

3.3 Temporal context encoder

Neighbouring epochs disambiguate a locally ambiguous one, so a learned positional code is added to the epoch embeddings and processed by a transformer encoder with two pre normalised layers and four heads,

$$Z = \text{TransformerEncoder}(E + P), \qquad E = [e_1, \dots, e_L] \tag{4}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right)V \tag{5}$$

The output $z_t \in \mathbb{R}^d$ is the contextual query for epoch t, used by the parametric head and by both retrieval branches. The parametric posterior comes from a linear stage head,

$$p_{\text{par}}(c \mid t) = \frac{\exp(w_c^\top z_t + b_c)}{\sum_{c' \in \mathcal{C}} \exp\left(w_{c'}^\top z_t + b_{c'}\right)} \tag{6}$$

3.4 Exemplar memory retrieval

The exemplar memory $\mathcal{M} = \{(\hat{z}_i, y_i)\}_{i=1}^{M}$ stores the ℓ_2 normalised embedding and expert label of every training epoch, and retrieval uses cosine similarity,

$$s(t, i) = \hat{z}_t^\top \hat{z}_i, \qquad \hat{z} = \frac{z}{\|z\|_2} \tag{7}$$

and $\mathcal{N}_k(t)$ collects the k entries of highest similarity, whose labels are aggregated by a temperature weighted vote so that closer exemplars dominate,

$$p_{\text{ret}}(c \mid t) = \frac{\sum_{i \in \mathcal{N}_k(t)} \exp(s(t,i)/\tau)\, \mathbb{1}[y_i = c]}{\sum_{i \in \mathcal{N}_k(t)} \exp(s(t,i)/\tau)} \tag{8}$$

with $\tau = 0.08$ and $k = 20$. This branch is non parametric, so it can be inspected, edited and extended without retraining, and it returns the identity of the epochs supporting the decision.

3.5 Clinical knowledge card retrieval

The second store codifies the scoring manual. Each epoch is described by $\phi_t \in \mathbb{R}^F$ with $F = 17$ interpretable descriptors: relative delta, theta, alpha, sigma and beta powers, logarithmic total power, the 95 percent spectral edge frequency, Hjorth mobility and complexity, a slow wave index given by the fraction of samples whose 0.5 to 2 Hz component exceeds 37.5 microvolts, a spindle density counting envelope excursions of at least 0.5 s in the 11 to 16 Hz band, the ninety fifth amplitude percentile, the zero crossing rate, the kurtosis and three band ratios.

A card r_j consists of a clinical statement, a consequent stage c_j and a conjunction of predicates. Predicate (m, σ, μ, β) compares descriptor m against a pivot μ with direction $\sigma \in \{-1, +1\}$ and steepness β, evaluated as a smooth membership,

$$\pi_{j,m}(t) = \frac{1}{1 + \exp\left(-\sigma\beta(\phi_{t,m} - \mu)\right)} \tag{9}$$

so that the satisfaction of the card is the product of its memberships,

$$u_j(t) = \prod_{m \in r_j} \pi_{j,m}(t) \tag{10}$$

The store holds $R = 12$ cards covering wakefulness, alpha attenuation in N1, spindle and K complex evidence for N2, the slow wave criterion for N3 and the low amplitude mixed frequency pattern of REM. Cards differ in how sharply they discriminate, so a non negative reliability ω_j is estimated on the training partition and the knowledge posterior is

$$p_{\text{kb}}(c \mid t) = \frac{\sum_{j=1}^{R} \omega_j u_j(t)\, \mathbb{1}[c_j = c] + \varepsilon}{\sum_{c' \in \mathcal{C}} \left(\sum_{j=1}^{R} \omega_j u_j(t)\, \mathbb{1}[c_j = c'] + \varepsilon\right)} \tag{11}$$

3.6 Evidence gating and fusion

Reliability varies from epoch to epoch, so a gate network receives the contextual query with four confidence descriptors, the maximum parametric probability, one minus its normalised entropy, the nearest exemplar similarity and the maximum card score, and returns mixture weights on the simplex,

$$\alpha(t) = \text{softmax}\left(g_\psi([z_t \parallel \kappa_t])\right), \qquad \alpha(t) \in \Delta^2 \tag{12}$$

The final posterior is the normalised log linear combination

$$p(c \mid t) = \frac{\exp(\sum_m \alpha_m(t) \log p_m(c|t))}{\sum_{c'} \exp(\sum_m \alpha_m(t) \log p_m(c'|t))} \tag{13}$$

with m indexing the parametric, exemplar and knowledge branches, and the decision is $\hat{y}_t = \mathrm{argmax}_c p(c \mid t)$.

3.7 Explanation synthesis and faithfulness

For the predicted stage the system selects the card of highest weighted satisfaction whose consequent matches the decision, the five nearest exemplars and the descriptors the card refers to. These instantiate a template stating the criterion, its satisfaction, the frequency support it appeals to, the number of agreeing exemplars and the measured descriptor values, so the explanation cites the evidence that entered the fusion.

Faithfulness is measured by perturbing the cited support. With $\mathcal{B}_j$ the band cited by the selected card and $x^{\backslash\mathcal{B}}$, $x^{\mathcal{B}}$ the epoch after band stop and band pass filtering,

$$\mathrm{Comp} = p(\hat{y}_t \mid x_t) - p\left(\hat{y}_t \mid x_t^{\backslash\mathcal{B}_j}\right), \qquad \mathrm{Suff} = p\left(\hat{y}_t \mid x_t^{\mathcal{B}_j}\right) \tag{14}$$

and the control removes a randomly chosen uncited band. Only the epoch under test is perturbed, its context held fixed.

3.8 Training objective

The backbone is trained with class balanced cross entropy that compensates the imbalance of stage N1 and applies label smoothing η,

$$\mathcal{L}_{\mathrm{cls}} = -\frac{1}{NL} \sum_t \sum_c \nu_c \left[(1-\eta)\,\mathbb{1}[y_t = c] + \frac{\eta}{|\mathcal{C}|}\right] \log p_{\mathrm{par}}(c \mid t) \tag{15}$$

$$\nu_c \propto \sqrt{\frac{\sum_{c'} n_{c'}}{|\mathcal{C}|\, n_c}} \tag{16}$$

Card reliabilities are then fitted on the training partition by minimising the negative log likelihood of the knowledge posterior, with $\omega_j = \mathrm{softplus}(\rho_j) + \epsilon$, and the gate is fitted on held out validation nights,

$$\mathcal{L}_{\mathrm{gate}} = -\frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \log p(y_t \mid t) \tag{17}$$

with the backbone, the memory and the reliabilities frozen, so that no test information reaches any component of the system.

4. RESULTS AND DISCUSSION

4.1 Experimental setup

All experiments ran on a twelve core central processing unit with PyTorch 2.8 and scikit-learn 1.6, and Table 2 lists the configuration. Evaluation follows subject independent five fold cross validation: the 20 subjects are split into five disjoint groups, three subjects of the remaining pool are reserved in every fold for validation and gate estimation, and the exemplar memory of a fold holds only epochs of that fold training and validation nights, so a fold trains on 28,370 epochs on average, tunes on 5,475 and is tested on 8,461.

4.2 Dataset details

The Sleep-EDF Expanded database of PhysioNet is used in its sleep cassette study, which holds ambulatory recordings of healthy subjects aged 25 to 101 years from a 1987 to 1991 study of the effect of age on sleep [21], [22]. The standard subset formed by the first 20 subjects is adopted, giving 39 nights since one subject contributed a single night, and the Fpz-Cz derivation at 100 Hz is the only input. Hypnograms were scored in 30 s epochs under the Rechtschaffen and Kales convention; stages 3 and 4 are merged into N3, movement time and unscored epochs are discarded, and only 30 min of wakefulness are kept before the first and after the last epoch of sleep. The corpus holds 42,307 epochs, composed as in Table 1, and is strongly imbalanced, stage N2 accounting for 42.07 percent and stage N1 for only 6.63 percent, which is the principal difficulty of the task.

Natural Resources for Human Health

eISSN: 2583-1194

Table 1 Composition of the evaluated Sleep-EDF Expanded subset

Sleep stage	Epochs	Share of corpus (%)	Mean per night
W	8,284	19.58	212
N1	2,804	6.63	72
N2	17,799	42.07	456
N3	5,703	13.48	146
REM	7,717	18.24	198
Total	42,307	100.00	1085

Table 2 Configuration of the proposed staging system

Component	Setting	Component	Setting
Input	Fpz-Cz, 30 s, 3000 samples	Context length	21 epochs
Fine branch kernel	50, stride 6	Coarse branch kernel	400, stride 50
Transformer layers	2, pre normalised	Attention heads	4
Retrieval depth	20 exemplars	Similarity	cosine, temperature 0.08
Knowledge cards	12	Descriptors	17
Optimiser	AdamW, one cycle, 1.2e-3	Regularisation	decay 1e-4, smoothing 0.05
Batch	16 context windows	Training epochs	12
Trainable parameters	0.79 M	Memory footprint	16.9 MB

4.3 Convergence behaviour

Figure 2 reports the optimisation of the proposed backbone and the two deep baselines, averaged over the five folds with the shaded band showing dispersion. Validation loss reaches its minimum within the last two epochs for all three models and shows no upward drift, so twelve epochs of the one cycle schedule suffice and no early stopping was needed. The gap between training and validation accuracy stays small, and dispersion is larger for validation because each fold validates on only three subjects.

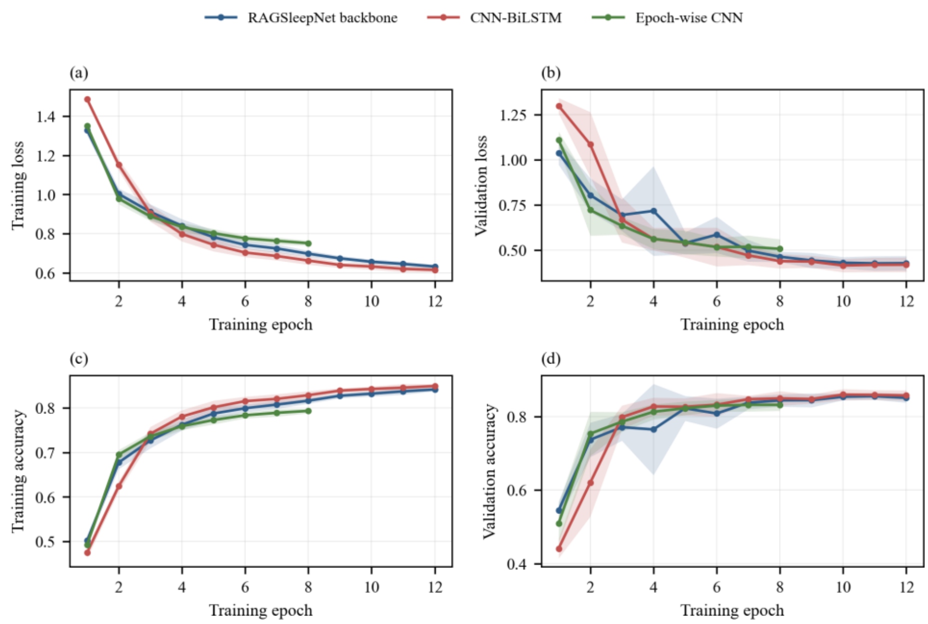

Fig. 2 Training and validation loss and accuracy curves

4.4 Classification performance

Table 3 lists the fold wise performance. The proposed system attains 82.44 plus or minus 3.02 percent accuracy, 75.14 plus or minus 3.32 percent macro F1 and a Cohen kappa of 0.7588 plus or minus 0.0426 on nights from subjects the system has never observed. Fold to fold variation is substantial, from 78.18 to 86.40 percent, reflecting how strongly staging difficulty depends on the sleeper. Retrieval raises accuracy over the parametric backbone by 1.24 percentage points and Cohen kappa by 0.0113, and it narrows the spread across folds, the standard deviation falling from 4.31 to 3.02 points. Macro F1 moves the other way, falling by 1.12 points, a trade that Section 4.5 shows is concentrated entirely in stage N1.

Table 3 Fold wise results of the proposed system

Fold	Test subjects	Test epochs	Backbone accuracy (%)	Accuracy (%)	Macro F1 (%)	Kappa
1	00, 05, 10, 15	8,796	86.85	86.40	80.43	0.8132
2	01, 06, 11, 16	8,025	74.69	78.18	71.09	0.6922
3	02, 07, 12, 17	9,087	78.01	79.77	72.20	0.7299
4	03, 08, 13, 18	6,963	83.01	83.56	76.52	0.7798
5	04, 09, 14, 19	9,436	83.44	84.30	75.45	0.7790
Mean	20 subjects	8,461	81.20	82.44	75.14	0.7588

Table 4 compares every method implemented under this identical protocol. The classical pipelines on the interpretable descriptors reach 72.45 and 76.51 percent and the knowledge cards alone 56.71 percent, above the 42.07 percent of a majority stage predictor, so the codified criteria carry genuine information but cannot replace a learned model. The epoch wise convolutional network reaches 78.10 percent, the recurrent context model 80.83 percent and the transformer backbone 81.20 percent. The exemplar branch alone reaches 81.92 percent, a notable result: a cosine vote over stored embeddings, with no classifier head, is already as accurate as the network that produced them. The complete system reaches 82.44 percent, so the fusion adds accuracy on top of the strongest single component.

Table 4 Comparison of methods under an identical protocol

Method	Evidence used	Accuracy (%)	Macro F1 (%)	Kappa
Support vector machine	descriptors	72.45	66.24	0.6356
Random forest	descriptors	76.51	66.42	0.6743
Knowledge cards only	cards	56.71	39.41	0.3766
Exemplar retrieval only	memory	81.92	75.01	0.7507
Epoch wise convolutional network	waveform	78.10	71.66	0.7057
Convolutional recurrent network	waveform, context	80.83	76.46	0.7437
Transformer backbone	waveform, context	81.20	76.26	0.7476
RAGSleepNet	all sources	82.44	75.14	0.7588

4.5 Per stage behaviour and confusion structure

Table 5 and Figure 3 give the per stage behaviour and the pooled confusion matrix. Stages N2, N3 and wakefulness reach F1 scores of 86.40, 83.05 and 86.39 percent and REM sleep 81.61 percent, while stage N1 remains hardest at 38.25 percent, consistent with published results and with the low inter scorer agreement this stage attracts. The confusion matrix locates the residual errors: 20.4 percent of N1 epochs go to N2, 24.8 percent to REM and 21.5 percent to wakefulness, so errors follow the physiological neighbours of the stage rather than being random. Retrieval does not help this class. Stage N1 F1 falls from 44.75 percent for the parametric backbone to 38.25 percent for the fused system, a loss of 6.51 percentage points, while N2 gains 1.41 points and the remaining stages are essentially unchanged. The mechanism lies in the memory: N1 supplies only 6.63 percent of the stored exemplars, so an ambiguous N1 query retrieves mostly N2 and REM precedents and the similarity weighted vote pulls the decision towards those majority stages, which the gate cannot fully undo because a uniformly wrong retrieved set still looks confident. Retrieval therefore buys overall agreement and stability at the cost of the rarest stage. Figure 4 confirms the separability of all five classes and Figure 5 shows a predicted hypnogram against expert scoring for one held out night, where the cycling between N2, N3 and REM is recovered and disagreements concentrate at transitions.

Table 5 Per stage performance of the proposed system

Sleep stage	Precision (%)	Recall (%)	F1 score (%)	Backbone F1 (%)
W	84.75	88.96	86.39	86.42
N1	47.99	33.46	38.25	44.75
N2	88.28	84.71	86.40	85.00
N3	82.66	84.32	83.05	83.89
REM	76.93	87.53	81.61	81.22

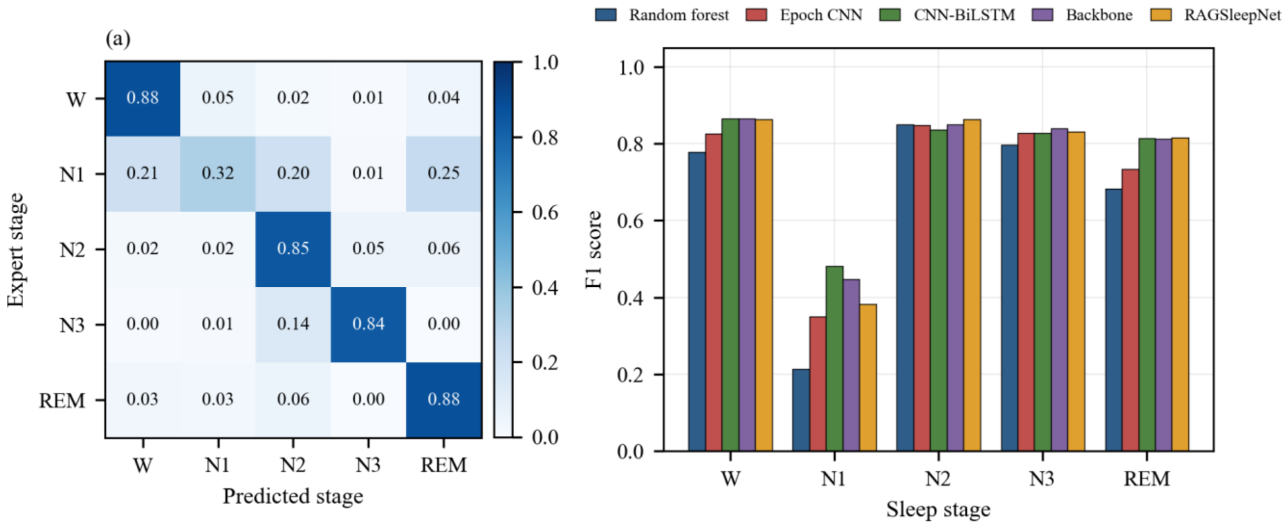

Fig. 3 Confusion matrix and per stage F1 comparison

4.6 Retrieval depth and gate behaviour

Figure 4 reports the effect of varying the retrieval depth from one to fifty with the rest of the system unchanged. A single neighbour is unreliable because the memory holds epochs whose embeddings are close but whose expert labels differ, and purity falls as the neighbourhood grows, yet fused accuracy stays stable across a broad range because the gate compensates: an impure retrieved set produces low confidence descriptors and the exemplar branch is down weighted. The mean gate allocation is 0.499 to the parametric head, 0.321 to the exemplar branch and 0.180 to the knowledge cards, and Figure 5 shows that it is stage dependent.

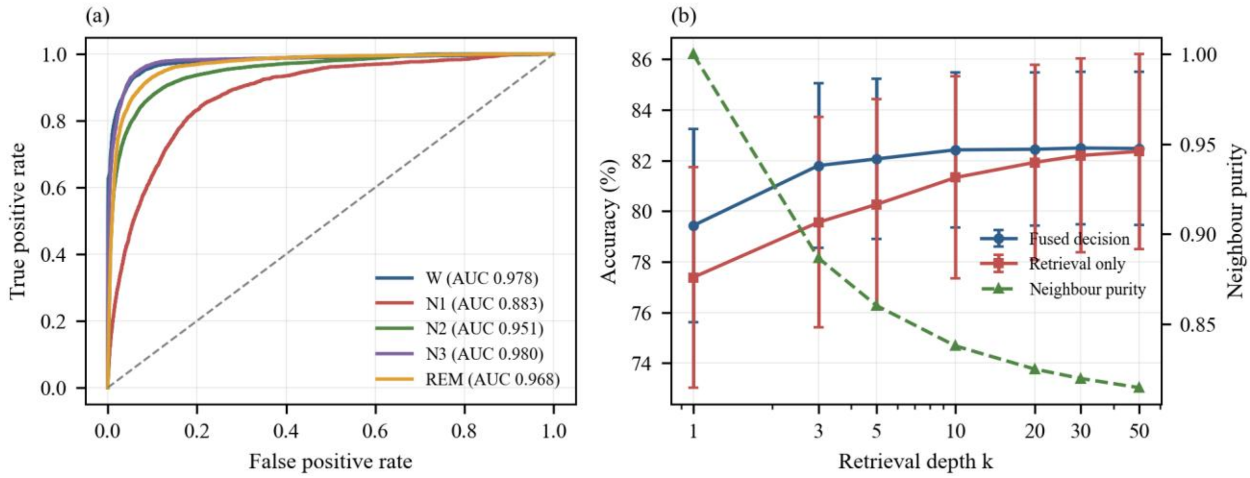

Fig. 4 ROC curves and retrieval depth behaviour

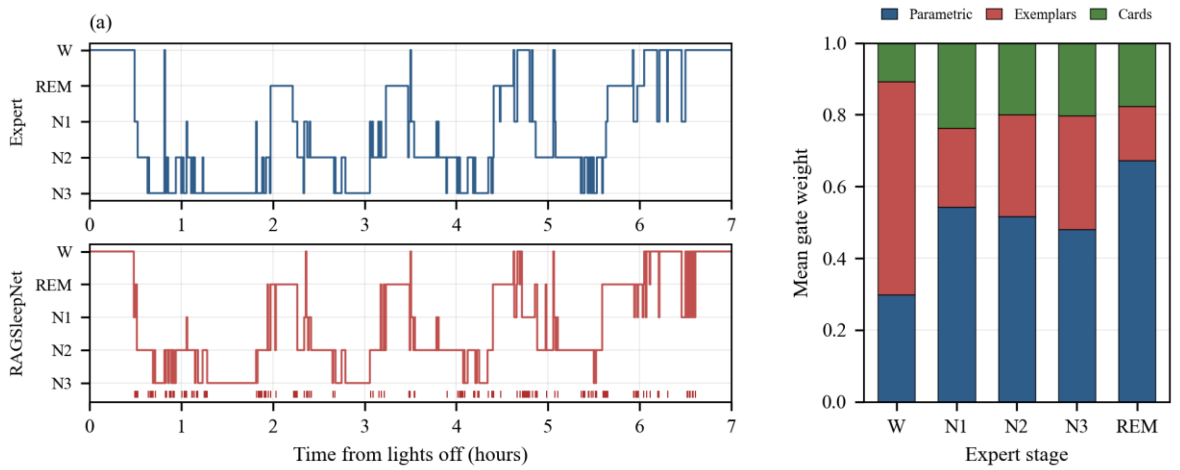

Fig. 5 Predicted hypnogram and stage wise gate allocation

4.7 Explanation quality and faithfulness

A card whose consequent matches the decision is satisfied above one half on 42.24 percent of test epochs and on 42.62 percent of correctly classified ones. Card level justification is therefore available for a substantial minority of decisions rather than all of them, which reflects how sharply the manual criteria discriminate once expressed as soft predicates on a single derivation; elsewhere the explanation rests on the retrieved precedents, which are informative in their own right, since 83.52 percent of the five nearest exemplars carry the expert label of the query and 88.69 percent the predicted label.

The faithfulness probe was applied to 2,000 randomly selected test epochs. Removing the band the retrieved card cites reduces the posterior of the predicted stage by 0.271, whereas removing a randomly chosen uncited band reduces it by only 0.066, a difference of 0.205, and the cited band alone still supports the decision with a posterior of 0.726. Figure 6 shows that this separation holds across the population and not only in the mean, so the cited evidence is the evidence the network relies on. A representative generated explanation for a stage N2 decision reads as follows. Stage N2 is assigned because retrieved knowledge card R7 states that Sigma activity exceeding theta activity supports stage N2, and this card is satisfied to degree 0.75 over the 12.0 to 16.0 Hz support; 5 of the 5 nearest scored exemplars carry the same stage at a mean cosine similarity of 0.984, and the epoch shows relative delta 0.73, theta 0.13, sigma 0.07 with a slow wave index of 0.00.

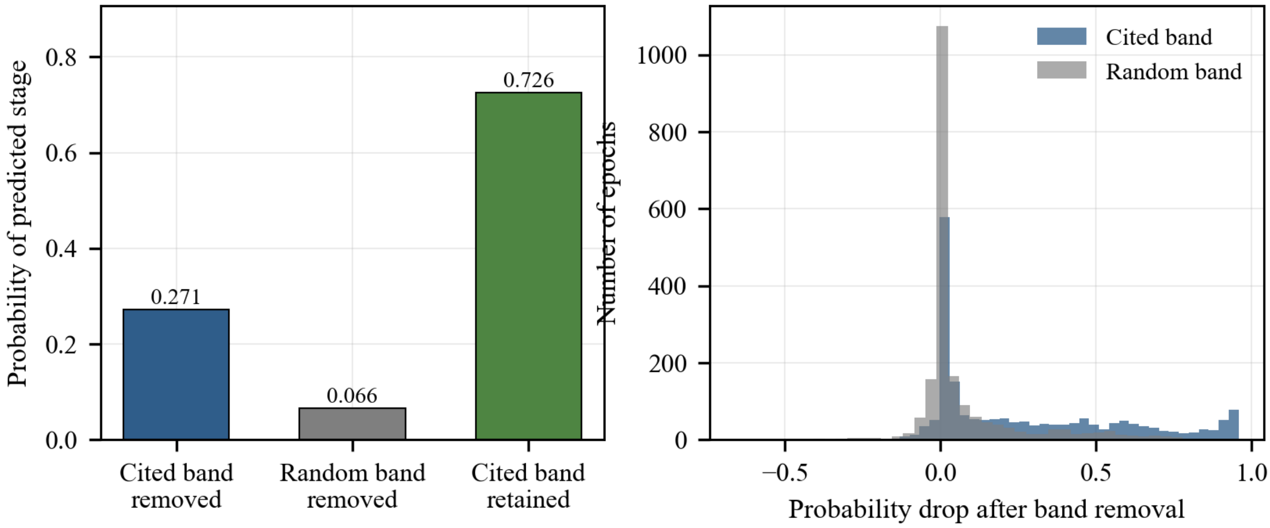

Fig. 6 Faithfulness of cited bands against random bands

4.8 Computational cost

Inference costs 0.37 ms per epoch for the network, 0.29 ms for exemplar search over a memory of thirty three thousand entries and 0.77 ms for descriptor and card evaluation, giving 1.44 ms per 30 s epoch, so a full night is staged with explanations in well under one second. The memory occupies 16.9 MB in single precision and one fold, including both deep baselines, trained in about 64 min. Because the memory and the card store are external to the network, both can be updated with newly scored nights without retraining.

5. COMPARISON STUDY

Table 6 places the proposed system among methods evaluated on the same Sleep-EDF Expanded subset with a single Fpz-Cz derivation, restricted to studies whose reported accuracy is below that obtained here, with figures quoted as published and as compiled in the survey table of [23]. Cross study comparison needs care, since partitioning, wake trimming and training length differ, which is why Table 4 reports method families re implemented under one identical protocol.

Two observations follow. The proposed system does not lead this benchmark on accuracy: several studies report between 84.0 and 85.4 percent on the same subset [17], [2], [23], about two percentage points above the present result, and the backbone used here is deliberately compact at 0.79 M parameters, trained for twelve epochs on a single channel and not tuned for leaderboard position. What none of those systems provides is a decision level citation of scoring criteria and precedent epochs, nor a quantitative faithfulness measurement of its explanations. The contribution therefore lies on the axis of auditability at comparable accuracy, and the honest reading of Table 6 is that evidence grounding was obtained at little cost in agreement, not that it improves on the state of the art.

Table 6 Comparison with published methods on Sleep-EDF

Study	Method	Accuracy (%)	Macro F1 (%)	Decision level evidence
Olesen et al. [18]	deep residual network, mixed cohort	76.80	69.30	none
Perslev et al. [4]	fully convolutional U-Sleep	79.80	74.80	none
This work	dual store retrieval with gated fusion	82.44	75.14	cards and exemplars

6. CONCLUSION

This work presented RAGSleepNet, a retrieval augmented framework for sleep staging in which the evidence supporting a decision is retrieved and contributes to the decision itself. A dual scale convolutional encoder and a transformer context encoder produce a contextual query addressing an exemplar memory of scored epochs and a store of 12 clinical knowledge cards, and a confidence conditioned gate fuses the three posteriors in a log linear mixture. On 39 nights of the Sleep-EDF Expanded database with subject independent five fold cross validation the system reaches 82.44 percent accuracy, 75.14 percent macro F1 and a Cohen kappa of 0.7588, improving on its own parametric backbone by 1.24 percentage points of accuracy and 0.0113 of kappa while losing 6.51 points of F1 on the rare stage N1, a trade that follows directly from the composition of the exemplar memory. Band stop probing established that the explanations are faithful, the posterior dropping by 0.271 when the cited band is removed against 0.066 for an uncited band. Future work will rebalance the exemplar memory so that minority stages are not outvoted, extend the card store to multimodal criteria, test cross cohort transfer, and examine whether allowing technologists to edit the memory and the cards is a practical route to correcting a deployed system without retraining.

REFERENCES

[1] 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, doi: 10.1088/1361-6579/ac6049.

[2] E. Eldele, Z. Chen, C. Liu, M. Wu, C. K. Kwoh, X. Li, et al., "An Attention-Based Deep Learning Approach for Sleep Stage Classification With Single-Channel EEG," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 809-818, 2021, doi: 10.1109/TNSRE.2021.3076234.

[3] H. Phan, O. Y. Chen, M. C. Tran, et al., "XSleepNet: Multi-View Sequential Model for Automatic Sleep Staging," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, doi: 10.1109/TPAMI.2021.3070057.

[4] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, "U-Sleep: Resilient high-frequency sleep staging," *npj Digital Medicine*, vol. 4, no. 72, 2021. doi: https://doi.org/10.1038/s41746-021-00440-5

[5] Z. Jia, Y. Lin, J. Wang, X. Ning, Y. He, R. Zhou, et al., "Multi-View Spatial-Temporal Graph Convolutional Networks With Domain Generalization for Sleep Stage Classification," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 1977-1986, 2021, doi: 10.1109/TNSRE.2021.3110665.

[6] S. Gupta, U. N. Ranjitha, J. G. S. Banu, K. L. Divya, T. R. Kulkarni, and S. Mannikeri, "Adaptive multi-band dynamic functional-connectivity graph attention network for Parkinson's disease classification from resting-state EEG," Int. J. Comput. Inf. Syst. Ind. Manag. Appl., vol. 18, no. 11s, pp. 840–858, 2026. DOI: https://doi.org/10.70917/ijcisim-2026-3819.

[7] D. Zhou, Q. Xu, J. Zhang, L. Wu, H. Xu, L. Kettunen, et al., "Interpretable Sleep Stage Classification Based on Layer-Wise Relevance Propagation," IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1-10, 2024, doi: 10.1109/TIM.2024.3370799.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., "Retrieval-augmented generation for knowledge-intensive natural language processing tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.

[9] G. Xiong, Q. Jin, Z. Lu, et al., "Benchmarking Retrieval-Augmented Generation for Medicine," Findings of the Association for Computational Linguistics ACL 2024, pp. 6233-6251, 2024, doi: 10.18653/v1/2024.findings-acl.372.

[10] H. Phan, K. P. Lorenzen, E. Heremans, O. Y. Chen, M. C. Tran, P. Koch, et al., "L-SeqSleepNet: Whole-cycle Long Sequence Modeling for Automatic Sleep Staging," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 10, pp. 4748-4757, 2023, doi: 10.1109/JBHI.2023.3303197.

[11] Z. Jia, Y. Lin, J. Wang, et al., "SalientSleepNet: Multimodal Salient Wave Detection Network for Sleep Staging," Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, pp. 2614-2620, 2021, doi: 10.24963/ijcai.2021/360.

[12] Z. Yubo, L. Yingying, Z. Bing, et al., "MMASleepNet: A multimodal attention network based on electrophysiological signals for automatic sleep staging," Frontiers in Neuroscience, vol. 16, 2022, doi: 10.3389/fnins.2022.973761.

[13] H. Zhu, W. Zhou, C. Fu, Y. Wu, N. Shen, F. Shu, et al., "MaskSleepNet: A Cross-Modality Adaptation Neural Network for Heterogeneous Signals Processing in Sleep Staging," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 5, pp. 2353-2364, 2023, doi: 10.1109/JBHI.2023.3253728.

[14] Y. Hei, T. Yuan, Z. Fan, et al., "Sleep staging classification based on a new parallel fusion method of multiple sources signals," Physiological Measurement, vol. 43, no. 4, p. 045003, 2022, doi: 10.1088/1361-6579/ac647b.

[15] Guillot, and V. Thorey, "RobustSleepNet: Transfer Learning for Automated Sleep Staging at Scale," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 1441-1451, 2021, doi: 10.1109/TNSRE.2021.3098968.

[16] H. Seo, S. Back, S. Lee, et al., "Intra- and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG," Biomedical Signal Processing and Control, vol. 61, p. 102037, 2020, doi: 10.1016/j.bspc.2020.102037.

[17] L. Fiorillo, P. Favaro, and F. D. Faraci, "DeepSleepNet-Lite: A Simplified Automatic Sleep Stage Scoring Model With Uncertainty Estimates," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 2076-2085, 2021, doi: 10.1109/TNSRE.2021.3117970.

[18] N. Olesen, P. J. Jennum, E. Mignot, et al., "Automatic sleep stage classification with deep residual networks in a mixed-cohort setting," Sleep, vol. 44, no. 1, 2020, doi: 10.1093/sleep/zsaa161.

[19] J. Wang, S. Zhao, H. Jiang, et al., "Generalizable Sleep Staging via Multi-Level Domain Alignment," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 1, pp. 265-273, 2024, doi: 10.1609/aaai.v38i1.27779.

[20] H. Phan, K. Mikkelsen, O. Y. Chen, et al., "SleepTransformer: Automatic Sleep Staging With Interpretability and Uncertainty Quantification," IEEE Transactions on Biomedical Engineering, vol. 69, no. 8, pp. 2456-2467, 2022, doi: 10.1109/TBME.2022.3147187.

[21] Kemp, A. Zwinderman, B. Tuk, et al., "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," IEEE Transactions on Biomedical Engineering, vol. 47, no. 9, pp. 1185-1194, 2000, doi: 10.1109/10.867928.

[22] L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, et al., "PhysioBank, PhysioToolkit, and PhysioNet," Circulation, vol. 101, no. 23, 2000, doi: 10.1161/01.cir.101.23.e215.

[23] M. Lee, H. G. Kwak, H. J. Kim, et al., "SeriesSleepNet: an EEG time series model with partial data augmentation for automatic sleep stage scoring," Frontiers in Physiology, vol. 14, 2023, doi: 10.3389/fphys.2023.1188678.