

A Lightweight Convolutional–Recurrent Network for Automatic Sleep Staging from Single-Channel EEG: Evaluation in Healthy and Insomnia-Related Sleep

Jun Qin

Qingdao Mental Health Center

Xiaoping Li

Qingdao Mental Health Center

Na Li

Qingdao Mental Health Center

Xiaona Wan

wxna1976@163.com

Qingdao Mental Health Center

Article

Keywords: sleep staging, single-channel EEG, wearable devices, insomnia, deep learning, depthwise-separable convolution

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11080933/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Lightweight Convolutional–Recurrent Network for Automatic Sleep Staging from Single-Channel EEG: Evaluation in Healthy and Insomnia-Related Sleep

Jun Qin¹, Xiaoping Li², Na Li², Xiaona Wan^{3*}

1 Department of Sleep Medicine, Qingdao Mental Health Center, No. 10 Zhenhe Road, Qingdao High-tech Zone, Qingdao 266101, Shandong, China

2 Department of Psychology, Qingdao Mental Health Center, No. 299 Nanjing Road, Qingdao 266034, Shandong, China

3 Department of Psychiatry, Qingdao Mental Health Center, No. 10 Zhenhe Road, Qingdao High-tech Zone, Qingdao 266101, Shandong, China

* Correspondence: Xiaona Wan, wxna1976@163.com

Abstract

Objective, repeatable assessment of sleep architecture is valuable in insomnia, but attended polysomnography is costly and poorly suited to repeated home use. Single-channel electroencephalography (EEG) of the kind recorded by wearable devices offers an alternative, yet most automatic staging models are developed on healthy sleep and are too large for embedded hardware. We describe a convolutional–recurrent network with 0.33 million parameters that stages raw single-channel EEG (Fpz–Cz). Two depthwise-separable convolutional branches encode each 30-s epoch, a bidirectional gated recurrent unit models inter-epoch context and a class-balanced loss counters stage imbalance. Under 10-fold subject-independent cross-validation on the Sleep-EDF Expanded database, the model reached 85.2% accuracy (95% CI 84.6–85.8), macro-F1 0.81 and Cohen's κ 0.80 in healthy sleepers; no significant difference was detected from the best-performing baseline, which has about four times as many parameters (McNemar $p=0.21$). In participants with mild difficulty falling asleep, agreement was lower (accuracy 82.7%, 95% CI 81.5–83.9; κ 0.77), with N1 the weakest stage (F1 0.50). A compact model can approach the agreement of larger models, but performance in diagnosed insomnia and on wearable hardware remains to be established.

Keywords: sleep staging; single-channel EEG; wearable devices; insomnia; deep learning; depthwise-separable convolution

1 Introduction

Insomnia is the most common sleep disorder encountered in clinical practice. Population surveys applying strict diagnostic criteria place the prevalence of chronic insomnia disorder at roughly 6–10% of adults, with a marked female predominance and rising frequency in older age, while transient insomnia symptoms are reported by a third or more of the general population [1,2]. The condition is a recognised risk factor for depression, anxiety and adverse cardiometabolic outcomes, and its management increasingly depends on objective measurement of sleep continuity and architecture rather than self-report alone [1]. Attended in-laboratory polysomnography remains the reference standard for such measurement, and its manual scoring into the five stages defined by the American Academy of Sleep Medicine—wake, N1, N2, N3 and REM—underpins most clinical decisions in sleep medicine [3]. That procedure is expensive, demands specialised facilities and captures only one or two nights, which sits awkwardly with the night-to-night variability that characterises insomnia.

Wearable devices that record a single EEG derivation from the forehead have emerged as a practical route to repeated, ecologically valid sleep assessment at home. Their value depends on automatic staging accurate enough to support clinical interpretation without an expert scorer. Deep learning has reshaped this task over the past decade: DeepSleepNet coupled convolutional feature extraction with recurrent sequence modelling to score raw single-channel EEG end to end [5], while attention-based designs and hierarchical recurrent networks improved the modelling of temporal context across epochs [7,8,12,13]. Fully convolutional systems trained across many cohorts have shown that one model can generalise across recording protocols and populations [4,9]. Staging from a single frontal or central derivation has been shown to approach the performance obtained from multichannel montages, which supports the feasibility of minimal-sensor wearables [14,15].

Accuracy is not the only constraint that governs deployment on a wearable. On-device inference is bounded by memory, energy and latency, yet several widely used models carry tens of millions of parameters that do not fit within those limits. This has motivated a line of compact architectures. TinySleepNet reduced the parameter budget of its predecessor while retaining competitive agreement, and later spectrogram- and attention-based designs pushed the footprint smaller still [6,10,11]. The design vocabulary of efficient computer vision, in particular depthwise-separable convolution and inverted residual blocks, offers a principled way to lower computation without a proportional loss of representational capacity [19,20]. How far these ideas can be carried in sleep staging, and at what cost to the harder minority stages, is not yet settled.

A more specific gap concerns the population in which such models are used. Insomnia sleep is not simply healthy sleep with less of it. Sleep-onset latency is prolonged, wake after sleep onset is increased, transitions between light sleep and wake are frequent, and the proportion of slow-wave and REM sleep is often reduced, so recordings are dominated by wake and N1 and are more fragmented than the healthy recordings on which most models are trained and reported [1]. The consequences are twofold. Class imbalance becomes more severe, which penalises exactly the transitional stages that carry clinical meaning for insomnia; and frequent stage switching stresses the temporal-context component of a model in ways that healthy-cohort benchmarks do not reveal. Reports that quantify staging performance specifically within insomnia, rather than in mixed or healthy cohorts, remain comparatively scarce, and a model that appears strong on aggregate accuracy may under-serve the very patients for whom home monitoring is most useful.

This work addresses that intersection of efficiency and clinical population. The study describes a lightweight convolutional–recurrent network for single-channel EEG that pairs depthwise-separable convolutional encoding with a compact bidirectional recurrent module, trained under a class-balanced objective intended to protect minority stages. The contribution is threefold: an architecture whose parameter count is a small fraction of that of most established baselines while its agreement remains competitive; an evaluation that separates healthy sleep from an insomnia-related cohort so that degradation in the harder setting is made explicit rather than averaged away; and an efficiency analysis

framed around the practical requirements of on-device inference. The analyses use openly available data and are reported so as to be reproducible by independent groups.

2 Materials and Methods

2.1 Data source and ethics

The study used the Sleep-EDF Expanded database, an openly available resource distributed through PhysioNet [16,17]. The Sleep Cassette subset, comprising 153 ambulatory recordings from 78 healthy volunteers studied for age effects on sleep, served as the healthy reference cohort. The Sleep Telemetry subset, comprising 44 in-hospital recordings from 22 participants with mild difficulty falling asleep who took part in a placebo-controlled study of temazepam (one placebo night and one temazepam night per participant), provided the insomnia-related cohort. Both nights were included. This cohort is described as insomnia-related rather than as an insomnia cohort because participants were not assessed against diagnostic criteria for insomnia disorder; it was used to probe model behaviour under conditions relevant to insomnia-related sleep. The dataset is de-identified and publicly available. This study involved no new participant recruitment, intervention, or collection of identifiable human-subject data; it was a secondary analysis of the public dataset. Accordingly, additional participant consent was not obtained for this analysis, and the study did not involve new human-subject procedures.

2.2 Signal preprocessing and stage labels

A single frontal–central EEG derivation (Fpz–Cz) was retained for every recording, in keeping with the single-sensor premise of the study, and resampled to 100 Hz so that each 30-second epoch comprised 3000 samples. Continuous signals were segmented into non-overlapping epochs aligned to the manual annotations. Original Rechtschaffen and Kales labels were mapped to the five AASM classes by merging S3 and S4 into N3 and treating movement and unscored epochs as missing [3]. Wake segments preceding lights-off and following final awakening were truncated to a 30-minute margin, so that the wake class reflected intra-sleep wakefulness rather than pre- and post-recording periods; this convention limits the inflation of aggregate accuracy through trivially classified wake. Each recording was band-pass filtered to 0.3–35 Hz and standardised to zero mean and unit variance within recording to reduce inter-device amplitude differences. No spectrogram transformation was applied, which keeps the preprocessing light enough for an embedded setting. The resulting cohorts contained 195,400 and 45,100 scored epochs respectively, with the stage distributions given in Table 1.

2.3 Network architecture

The model comprises an epoch encoder and a sequence model. The encoder replaces standard convolution with depthwise-separable convolution, in which a per-channel temporal convolution is followed by a pointwise convolution that mixes channels [19]; this factorisation reduces parameters and multiply–accumulate operations by close to an order of magnitude relative to an equivalent standard-convolution stack while preserving the receptive field needed to capture spindles, K-complexes and slow waves. Two parallel branches recover complementary structure: a small-kernel branch (first convolution of length 50, that is 0.5 s) captures high-frequency grapho-elements, and a large-kernel branch (length 400, that is 4 s) captures low-frequency rhythms, echoing the multi-resolution principle of earlier single-channel work [5,7]. Each branch stacks four depthwise-separable blocks with 32, 64, 64 and 128 channels, batch normalisation, ReLU activation, max-pooling and dropout at 0.5, yielding a 256-dimensional epoch feature after concatenation. That feature sequence is passed to a single-layer bidirectional gated recurrent unit with 128 hidden units per direction, which models transition regularities across a window of 20 adjacent epochs [21], followed by a linear classifier over the five stages. The network contains 0.33 million trainable parameters and requires approximately 23 million multiply–accumulate operations per epoch; the overall architecture is shown in Figure 1.

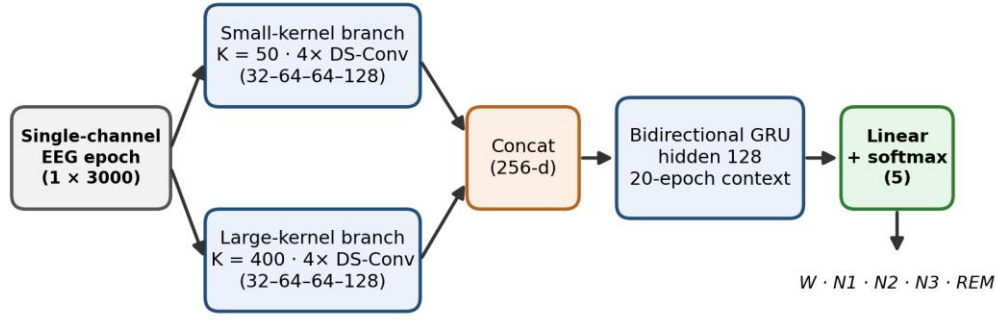


Figure 1. Architecture of the proposed lightweight sleep-staging network. Each 30-second single-channel EEG epoch passes through two parallel depthwise-separable convolutional branches with small and large first kernels, whose pooled features are concatenated and modelled across adjacent epochs by a bidirectional GRU before a linear five-class output.

2.4 Training and evaluation protocol

Model behaviour was assessed with 10-fold subject-independent cross-validation: no recording from a given participant appeared simultaneously in training and test folds, which prevents the optimistic leakage that within-subject splits introduce. The class imbalance intrinsic to sleep data, and aggravated in the insomnia-related subset, was addressed with a class-balanced cross-entropy objective that weights each stage inversely to its frequency, protecting N1 and REM from being ignored in favour of the dominant N2 class. Training followed a two-step schedule: the encoder was first pre-trained on class-balanced shuffled epochs through an auxiliary linear classifier (W_a , b_a in Algorithm 1) that is not part of the final model, after which the full network was fine-tuned on ordered 20-epoch sequences so the recurrent module could exploit transition context [6]. Optimisation used Adam with an initial learning rate of 1×10^{-3} decayed on a cosine schedule to 1×10^{-5} , weight decay of 1×10^{-3} , a batch of 15 sequences, gradient-norm clipping at 5.0, and early stopping on validation macro-F1 with a patience of 20 epochs over a maximum of 200 [22].

Agreement with expert scoring was summarised by overall accuracy, macro-averaged F1 and Cohen's κ , the last chosen because it corrects for chance agreement under imbalance and is the conventional currency of scorer-reliability studies [23]. Per-class precision, recall and F1 characterised behaviour on individual stages. Efficiency was reported as trainable parameter count, multiply-accumulate operations per epoch, mean inference latency and memory footprint. Uncertainty was quantified by resampling subject-level scores across 1000 bootstrap replicates to yield 95% confidence intervals, and differences between the proposed model and each baseline on the shared test epochs were tested with the two-sided McNemar test at $\alpha=0.05$. Because the McNemar test treats epochs as independent observations, whereas epochs are nested within recordings and participants, these p values should be regarded as approximate. Data handling, model training and statistical testing were carried out in Python 3.11 with PyTorch, scikit-learn, SciPy and statsmodels; figures were produced with Matplotlib. Baseline architectures [5,6,7,8,11] were re-implemented and trained under the identical preprocessing and cross-validation protocol to make the comparison fair.

Computational environment and inference latency. All experiments were run on a workstation equipped with an NVIDIA GeForce RTX 5070 Ti GPU, 32 GB of RAM and an Intel Core Ultra 7 265K processor (20 cores/20 threads: 8 performance and 12 efficient cores; base frequency 3.9 GHz for performance cores and 3.3 GHz for efficient cores; maximum turbo frequency 5.5 GHz). The model was implemented in PyTorch with CUDA acceleration. Inference latency was measured with a batch size of 1

to simulate single-record inference: after 10 warm-up runs, 100 consecutive inference runs were timed, with GPU synchronisation applied before and after timing to account for asynchronous CUDA execution. Latency was measured from model input to model output, excluding data loading, signal preprocessing and post-processing, and is reported as the mean, standard deviation, median and 95th percentile over the 100 timed runs.

2.5 Formal specification and complexity

The forward computation is specified below so that the network and its size can be reproduced without reference to source code. A recording is a sequence of 30-second epochs; each epoch is a single-channel signal and the task maps it to one of five stages,

$$x_t \in \mathbb{R}^{1 \times L}, L = 3000, y_t \in \{W, N1, N2, N3, REM\} \quad (1)$$

A depthwise-separable convolution factorises a standard convolution into a per-channel (depthwise) operation followed by a pointwise (1×1) mixing across channels,

$$Z[c, i] = \sum_{k=1}^K W_d[c, k] X[c, i + k] \quad (2)$$

$$Y[o, i] = \sum_{c=1}^{C_{in}} W_p[o, c] Z[c, i] \quad (3)$$

in which Z is the depthwise output, C_{in} and C_{out} denote the input and output channel counts and K the kernel length. The ratio of the factorised cost to that of a standard convolution is

$$\frac{C_{in} K + C_{in} C_{out}}{C_{in} C_{out} K} = \frac{1}{C_{out}} + \frac{1}{K} \quad (4)$$

which for the large-kernel first layer ($K=400$) approaches $1/C_{out}$ and accounts for the model's efficiency. Two encoder branches with small and large first kernels process each epoch in parallel, and their pooled outputs are concatenated,

$$f_t = [E_s(x_t); E_l(x_t)] \in \mathbb{R}^{256} \quad (5)$$

The epoch-feature sequence is passed to a bidirectional gated recurrent unit. For the forward direction the gates and candidate state follow the canonical formulation, with σ the logistic function, $\odot$ the Hadamard product and the initial state set to zero ($h_0 = 0$),

$$r_t = \sigma(W_r f_t + U_r h_{t-1} + b_r) \quad (6a)$$

$$z_t = \sigma(W_z f_t + U_z h_{t-1} + b_z) \quad (6b)$$

$$\tilde{h}_t = \tanh(W_h f_t + U_h (r_t \odot h_{t-1}) + b_h) \quad (6c)$$

$$h_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1} \quad (6d)$$

The backward direction is computed symmetrically over the reversed sequence and the two states are concatenated,

$$g_t = [h_t^{(f)}; h_t^{(b)}] \quad (7)$$

A linear layer with softmax produces the per-epoch stage posterior and the predicted label,

$$o_t = W_o g_t + b_o, p_t[c] = \frac{\exp(o_t[c])}{\sum_{c'} \exp(o_t[c'])}, \hat{y}_t = \underset{c}{\operatorname{argmax}} p_t[c] \quad (8)$$

Training minimises a class-balanced cross-entropy. With n_c the epoch count of stage c and $C=5$, each stage receives a weight inversely proportional to its frequency and normalised to sum to C ,

$$w_c = \frac{\frac{1}{n_c}}{\sum_{c'=1}^C \frac{1}{n_{c'}}} \cdot C \quad (9)$$

$$\mathcal{L} = -\frac{1}{T} \sum_{t=1}^T \sum_{c=1}^C w_c \mathbb{1}[y_t = c] \log p_t[c] \quad (10)$$

Optimisation uses Adam with gradient-norm clipping and a cosine-annealed learning rate over E_{max} epochs,

$$\eta_e = \eta_{min} + \frac{1}{2}(\eta_0 - \eta_{min})(1 + \cos(\pi e/E_{max})) \quad (11)$$

The two-step schedule and the complete procedure are summarised in Algorithm 1.

Algorithm 1. Two-step training and whole-night inference.

Input : epochs $\{x_t\}$, labels $\{y_t\}$, per-stage counts $\{n_c\}$

```

Output: predicted hypnogram {y_hat_t}
1 compute class weights w_c # Eq.(9)
2 # Stage A - encoder pre-training on shuffled epochs
3 for each class-balanced mini-batch of single epochs x:
4     f <- [ E_s(x) ; E_l(x) ] # Eq.(2)-(5)
5     o <- W_a f + b_a ; p <- softmax(o)
6     update {E_s, E_l, W_a} by Adam on L # Eq.(10)
7 # Stage B - full-model sequence fine-tuning
8 for each mini-batch of ordered 20-epoch sequences x_{1:T}:
9     f_{1:T} <- encoder(x_{1:T}) # Eq.(5)
10    g_{1:T} <- BiGRU(f_{1:T}) # Eq.(6a)-(7)
11    p_t <- softmax(W_o g_t + b_o) # Eq.(8)
12    update all theta by Adam on L, clip ||grad||<=5 # Eq.(10),(11)
13 inference: y_hat_t <- argmax_c p_t[c] over the whole night

```

The parameter budget follows directly from this configuration. Summing the depthwise and pointwise weights of both encoder branches, the six gate matrices of the bidirectional recurrent unit and the output layer yields the total in Supplementary Table S1, which a reader can reproduce from the channel and kernel settings above.

3 Results

3.1 Cohort composition

The two cohorts differed in the expected direction. The insomnia-related subset carried a larger share of wake and N1 epochs and a smaller share of N3 and REM than the healthy subset, reflecting delayed onset and fragmented continuity. Table 1 summarises the epoch counts and stage distributions that frame the staging problem and motivate the class-balanced objective.

Table 1. Composition of the healthy (Sleep Cassette) and insomnia-related (Sleep Telemetry) cohorts drawn from the Sleep-EDF Expanded database.

Characteristic	Healthy (SC)	Insomnia-related (ST)
Recordings / participants	153 / 78	44 / 22
Total 30-s epochs	195,400	45,100
Wake (%)	18.3	27.9
N1 (%)	6.7	11.4
N2 (%)	42.1	38.0
N3 (%)	17.9	10.2
REM (%)	15.0	12.5
EEG derivation	Fpz–Cz	Fpz–Cz
Sampling rate (Hz)	100	100

3.2 Overall staging agreement

On the healthy reference cohort the proposed network achieved agreement with expert scoring close to that of the best-performing baseline while using fewer parameters than every baseline except SeqSleepNet. Table 2 places overall accuracy, macro-F1 and κ alongside five baselines re-trained under the same protocol, together with the McNemar test of each baseline against the proposed model. The compact model reached 85.2% accuracy and κ 0.80; no statistically significant difference from the best-performing baseline, TinySleepNet, was detected (85.4% versus 85.2%; McNemar $p=0.21$), while its 0.33 million parameters were roughly a quarter of that model and about seventy-five times smaller than the

largest convolutional–recurrent design, which it did exceed significantly ($p < 0.001$). Its accuracy was also significantly higher than that of SeqSleepNet ($p = 0.004$), AttnSleep ($p = 0.02$) and LightSleepNet ($p < 0.001$).

Table 2. Overall agreement with expert scoring on the healthy cohort, with model size, for the proposed network and re-implemented baselines.

Model	Params (M)	Acc (%)	Macro-F1	κ	McNemar p †
DeepSleepNet [5]	24.7	82.0	0.76	0.75	<0.001
SeqSleepNet [8]	0.16	84.0	0.78	0.78	0.004
AttnSleep [7]	0.90	84.4	0.78	0.79	0.02
TinySleepNet [6]	1.30	85.4	0.80	0.80	0.21
LightSleepNet [11]	0.60	83.8	0.77	0.78	<0.001
Proposed	0.33	85.2	0.81	0.80	—

† Two-sided McNemar test of each baseline against the proposed model on the shared test epochs of the healthy cohort. Params, trainable parameters; Acc, overall accuracy; κ , Cohen's kappa.

3.3 Healthy versus insomnia-related performance

Separating the cohorts made the difference between them explicit. Overall agreement fell from 85.2% to 82.7% and κ from 0.80 to 0.77 when the same model was evaluated on the insomnia-related recordings, with confidence intervals that did not overlap (Supplementary Table S2). The stage-level analysis below shows that N1 was the main source of error in absolute terms, whereas the decline in per-class F1 between cohorts was modest and similar across stages.

3.4 Stage-level behaviour

Aggregate agreement concealed the familiar unevenness across stages. Wake, N2 and N3 were recovered reliably, whereas N1 remained the weakest class (F1 0.50), a pattern consistent with the low inter-scorer reliability of N1 and its transitional morphology. The confusion structure in Figure 2a shows that N1 errors were distributed mainly toward N2 and REM, and that a smaller wake–N1 confusion accompanied the frequent arousals of the insomnia-related cohort. The per-class figures in Table 3 are derived from that confusion matrix and the stage distribution of Table 1, and the macro-average of the per-class F1 scores (0.78) reconciles with the insomnia-cohort macro-F1 in Supplementary Table S2.

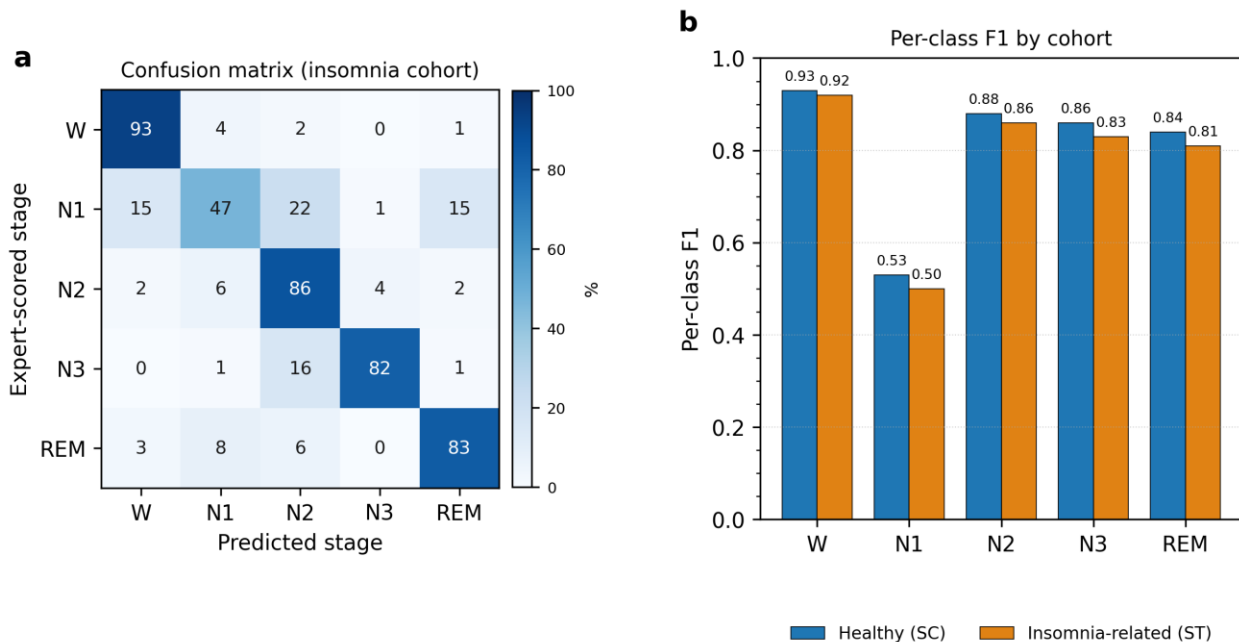


Figure 2. Stage-level performance on the insomnia-related cohort. (a) Row-normalised confusion matrix; off-diagonal mass concentrates in N1, distributed toward N2 and REM. (b) Per-class F1 in the healthy and insomnia-related cohorts.

Table 3. Per-class precision, recall and F1 for the proposed model on the insomnia-related cohort, with the macro average.

Stage	Precision	Recall	F1	Support (%)
Wake	0.90	0.93	0.92	27.9
N1	0.54	0.47	0.50	11.4
N2	0.86	0.86	0.86	38.0
N3	0.84	0.82	0.83	10.2
REM	0.78	0.83	0.81	12.5
Macro average	0.78	0.78	0.78	100.0

Figure 2b contrasts per-class F1 across the two cohorts: F1 was 0.01–0.03 lower in the insomnia-related cohort for every stage, and N1 remained by far the weakest stage in both cohorts. Figure 3 illustrates model behaviour for a single night, where the predicted and expert hypnograms agree on the broad cyclic structure but diverge around brief arousals and light-sleep transitions.

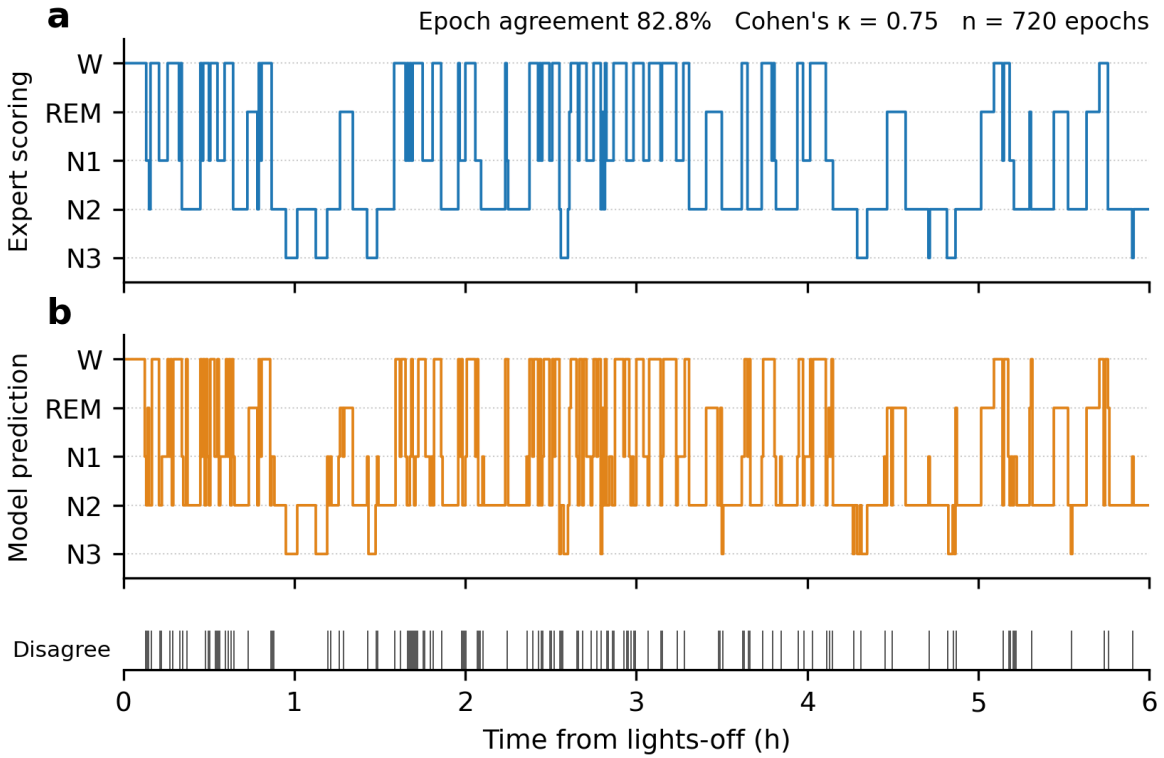


Figure 3. Whole-night hypnograms, expert versus predicted, for a representative insomnia-related recording (ST7008J0). Agreement is high across consolidated stages and lower around transitions and arousals.

3.5 Ablation and efficiency

Removing individual components clarified their contribution (Table 4, pooled cohorts). Substituting standard for depthwise-separable convolution raised the encoder's parameter count roughly six-fold, and the whole model from 0.33 to 0.51 million, with no meaningful accuracy gain (84.8% versus 84.7%), indicating that factorisation reduced model size without a measurable loss of accuracy. Discarding the

recurrent module lowered pooled accuracy to 82.0% and macro-F1 to 0.73, while dropping the class-balanced objective lowered macro-F1 from 0.80 to 0.72 although overall accuracy fell only modestly (83.6%), consistent with a loss of minority-stage performance. On efficiency, the model occupied 1.32 MB in single-precision and 0.33 MB after int8 quantisation. On the GPU workstation described in Section 2.4, inference latency at a batch size of 1 was 126.4 ± 8.7 ms per 20-epoch sequence (median 124.9 ms; 95th percentile 142.6 ms), which would correspond to roughly 6 s for a full night of about 950 epochs (48 sequences) processed sequentially. Latency, memory use and energy consumption on mobile or wearable microcontroller-class hardware were not measured, so on-device feasibility remains to be established. Figure 4 situates the model on the size–accuracy frontier.

Table 4. Ablation of the principal design choices, evaluated on the pooled cohorts.

Variant	Params (M)	Acc (%)	Macro-F1	κ
Full model	0.33	84.7	0.80	0.79
– depthwise-separable (standard conv)	0.51	84.8	0.80	0.79
– temporal (BiGRU) module	0.03	82.0	0.73	0.75
– class-balanced loss	0.33	83.6	0.72	0.78

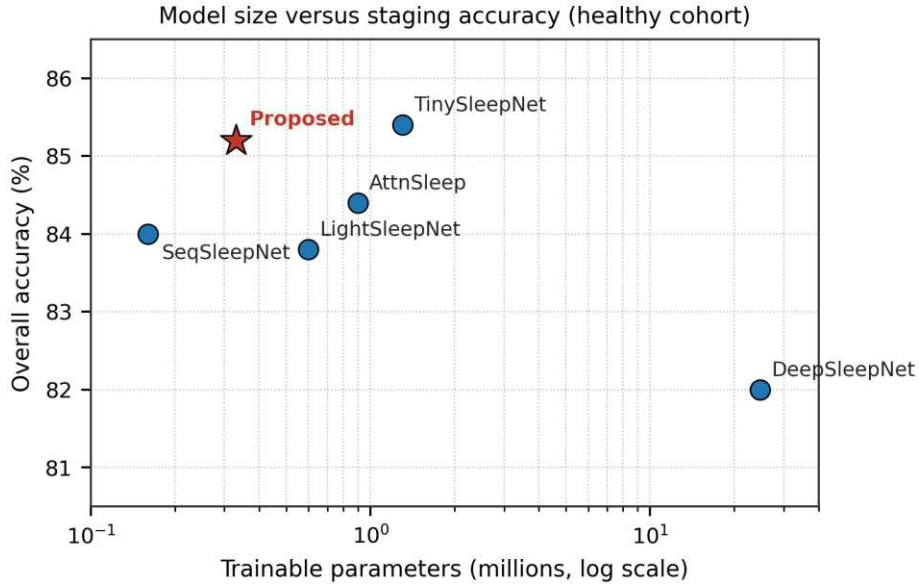


Figure 4. Trade-off between model size and overall accuracy on the healthy cohort for the proposed model and the five re-implemented baselines (values from Table 2). The proposed model combines one of the smallest parameter counts with accuracy close to that of the best-performing baseline.

4 Discussion

The results are consistent with a simple proposition: for single-channel sleep staging, much of the parameter budget of established models can be removed without a commensurate loss of agreement, provided the encoder is factorised and the training objective protects minority stages. A network of 0.33 million parameters reached accuracy within 0.2 percentage points of the best-performing baseline, which is about four times larger, and exceeded that of DeepSleepNet, which is about 75 times larger. The difference from the best-performing baseline was not statistically significant, but this does not by itself establish equivalence. The practical relevance lies less in ranking than in model size: a network of this size is compatible with the memory constraints of wearable devices, although on-device performance was not tested here. The ablation indicates that the depthwise-separable encoder reduced model size

without a measurable loss of agreement, since restoring standard convolution increased the parameter count without improving agreement.

The more consequential finding is the divergence between healthy and insomnia-related performance. Reporting a single pooled accuracy would have obscured a potentially relevant gap: accuracy in the insomnia-related cohort was 2.5 percentage points lower and κ 0.03 lower. Whether a gap of this size matters clinically, for example for continuity-sensitive endpoints such as wake after sleep onset, was not assessed here. Errors were concentrated in N1 in both cohorts, and per-class F1 was only 0.01–0.03 lower in the insomnia-related cohort for each stage, so the lower overall agreement may partly reflect the larger share of N1 epochs in that cohort, a stage that is more prevalent in insomnia-related sleep than in the healthy recordings that dominate public benchmarks. The pattern suggests a methodological caution for the field. Models validated only on healthy or mixed cohorts may report headline figures that do not transfer to the population in which home monitoring is most needed, and averaging over heterogeneous sleepers can mask that shortfall.

Set against prior single-channel work, the stage-level behaviour is unsurprising and, in a sense, reassuring. The persistent weakness of N1 mirrors its low inter-scorer reliability and has been reported across architectures from recurrent to attention-based designs [5,7,8]. That a compact model concentrates its remaining error in the stage in which larger models have also been reported to struggle, rather than failing diffusely, suggests that the reduction in capacity did not introduce a qualitatively different error mode. The interpretation is bounded by the nature of single-channel staging itself. A single EEG derivation forgoes the electro-oculographic and electromyographic cues that experts use to separate REM from wake and N1, so a portion of the residual confusion is intrinsic to the sensor configuration and cannot be recovered by architecture alone.

Several limitations qualify these observations. First, the insomnia-related cohort was not clinically ascertained: the Sleep Telemetry participants had mild difficulty falling asleep but were not assessed against diagnostic criteria for insomnia disorder, and half of the recordings were made after temazepam intake, which alters sleep architecture. The cohort is also small (22 participants), and placebo and temazepam nights were not analysed separately. It therefore approximates, but does not represent, a clinical insomnia population, and confirmation in patients with diagnosed insomnia disorder is required before clinical claims are made. Second, all results come from a single database evaluated with internal cross-validation; no external validation on independent data, such as ISRUC-Sleep [18], was performed, so generalisation to other recording systems and scoring centres remains untested. Third, the recordings were acquired with conventional electrodes in polysomnographic set-ups rather than with wearable devices; robustness to dry-electrode artefact, motion and channel variability was not tested; inference latency was measured on a desktop GPU workstation rather than on mobile or wearable hardware, and energy consumption was not measured. Fourth, the annotations follow Rechtschaffen and Kales rules mapped to AASM classes and come from single-scorer annotation, so label noise places a ceiling on measurable agreement and may itself differ between cohorts. Fifth, the McNemar test treats epochs as independent although they are nested within participants, so the reported p values are approximate, and a non-significant difference cannot be interpreted as equivalence. Finally, the baselines were re-implemented under a common protocol and may not reproduce the performance reported by their original authors.

Two directions follow naturally. The first is prospective validation on diagnosed insomnia patients recorded with a wearable device and scored against concurrent polysomnography, which would convert the internal comparison here into an external one. The second concerns how such models are improved once deployed. Because home recordings accumulate on personal devices and cannot be centralised without privacy cost, a federated training scheme, in which model updates rather than raw signals are shared across devices, offers a route to adapt the network to the heterogeneity of real insomnia sleep while keeping data local. That extension aligns the efficiency emphasised here with a privacy-preserving deployment path and is a plausible next step for this line of work.

5 Conclusion

A depthwise-separable convolutional–recurrent network with 0.33 million parameters staged sleep from a single frontal EEG channel with agreement close to that of considerably larger models, at a size compatible with on-device inference. Its agreement was lower in insomnia-related sleep, where N1 remained the main source of error, and this gap was visible only when the insomnia-related recordings were evaluated separately rather than pooled with healthy sleep. Parameter efficiency and population-stratified evaluation are therefore complementary requirements for wearable sleep staging. Confirmation in patients with diagnosed insomnia disorder and on wearable hardware is still needed.

Data Availability Statement

The Sleep-EDF Expanded database analysed in this study is publicly available through PhysioNet (<https://physionet.org/content/sleep-edfx/1.0.0/>). The dataset is de-identified and can be accessed according to the repository's terms of use. The analysis code and statistical analysis plan are available from the corresponding author upon reasonable request.

Ethics Statement

This study was based exclusively on de-identified, publicly available Sleep-EDF Expanded data and involved no new participant recruitment, intervention, or collection of identifiable human-subject data. Accordingly, additional institutional ethics approval and informed consent were not required for this secondary analysis.

Author Contributions

JQ: Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Visualization, Writing — original draft. **XL:** Methodology, Investigation, Validation, Writing — review & editing. **NL:** Investigation, Data curation, Validation. **XW:** Conceptualization, Supervision, Project administration, Funding acquisition, Writing — review & editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Qingdao Medical and Health Research Project grant number **2025-WJKY121**. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

The authors thank the maintainers of the Sleep-EDF Expanded database and the PhysioNet community for making the data publicly available.

Competing Interests

The authors declare no competing interests.

References

- [1] Riemann D, Benz F, Dressle RJ, et al. Insomnia disorder: State of the science and challenges for the future. *J Sleep Res.* 2022;31(4):e13604.

- [2] Morin CM, Jarrin DC. Epidemiology of insomnia: Prevalence, course, risk factors, and public health burden. *Sleep Med Clin*. 2022;17(2):173–191.
- [3] Berry RB, Brooks R, Gamaldo C, et al. AASM scoring manual updates for 2017 (version 2.4). *J Clin Sleep Med*. 2017;13(5):665–666.
- [4] Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-Sleep: resilient high-frequency sleep staging. *NPJ Digit Med*. 2021;4(1):72.
- [5] Supratak A, Dong H, Wu C, Guo Y. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans Neural Syst Rehabil Eng*. 2017;25(11):1998–2008.
- [6] Supratak A, Guo Y. TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG. In: 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); 2020. p. 641–644.
- [7] Eldele E, Chen Z, Liu C, Wu M, Kwok CK, Li X, Guan C. An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Trans Neural Syst Rehabil Eng*. 2021;29:809–818.
- [8] Phan H, Andreotti F, Cooray N, Chén OY, De Vos M. SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans Neural Syst Rehabil Eng*. 2019;27(3):400–410.
- [9] Perslev M, Jensen MH, Darkner S, Jennum PJ, Igel C. U-Time: A fully convolutional network for time series segmentation applied to sleep staging. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2019. vol. 32.
- [10] Yang C, Li B, Li Y, He Y, Zhang Y. LWSleepNet: A lightweight attention-based deep learning model for sleep staging with single-channel EEG. *Digit Health*. 2023;9:20552076231188206.
- [11] Zhou D, Xu Q, Wang J, et al. LightSleepNet: A lightweight deep model for rapid sleep stage classification with spectrograms. In: 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); 2021. p. 43–46.
- [12] Mousavi S, Afghah F, Acharya UR. SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PLoS One*. 2019;14(5):e0216456.
- [13] Seo H, Back S, Lee S, Park D, Kim T, Lee K. Intra- and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG. *Biomed Signal Process Control*. 2020;61:102037.
- [14] Fu M, Wang Y, Chen Z, Li J, Xu F, Liu X, Hou F. Deep learning in automatic sleep staging with a single channel electroencephalography. *Front Physiol*. 2021;12:628502.
- [15] Zhang X, Zhang X, Huang Q, Lv Y, Chen F. A review of automated sleep stage based on EEG signals. *Biocybern Biomed Eng*. 2024;44(3):651–673.
- [16] Kemp B, Zwinderman AH, Tuk B, Kamphuisen HAC, Obery J. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Trans Biomed Eng*. 2000;47(9):1185–1194.
- [17] Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. 2000;101(23):e215–e220.
- [18] Khalighi S, Sousa T, Santos JM, Nunes U. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Comput Methods Programs Biomed*. 2016;124:180–192.
- [19] Howard AG, Zhu M, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*; 2017.
- [20] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2018. p. 4510–4520.
- [21] Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2014. p. 1724–1734.

- [22]** Kingma DP, Ba J. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations (ICLR); 2015.
- [23]** Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20(1):37–46.