SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Information System Exploration and Research

https://joiser.shmpublisher.com

p-ISSN 2964-1160 | e-ISSN 2963-6361

Approaching the Human Ceiling in Sleep Staging: A Controlled Comparison of Feature-Engineered and Deep Architectures

Rand Nabeel Dawood[1*]
[1]*General Directorate of Public, Ministry of Education, Iraq*

DOI: https://doi.org/10.52465/joiser.v4i3.20
Received 30 May 2026; Accepted 16 July 2026; Available online 21 July 2026

Article Info

Keywords:
Hypnogram;
EEG;
EOG;
Sleeping stages;
Deep learning

Abstract

Polysomnography (PSG) is the clinical gold standard for sleep assessment, yet manual epoch-by-epoch scoring is labor-intensive and subject to substantial inter-rater variability, which limits its use in large-scale, home-based, and routine clinical practice and makes reliable automated scoring an urgent need. This study suggests and compares an automated sleep-staging model based on two different supervised pipelines: a conventional feature-engineered Random Forest and a state-of-the-art end-to-end Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) network. Based on the recordings of the Sleep-EDF Expanded database, we preprocessed the multichannel signals (EEG and EOG) into 30-second epochs, which were classified into five stages (Wake, N1, N2, N3 and REM). Our findings show that the feature-engineered baseline is very competitive with 95.49% accuracy, but the CNN-LSTM model performs better (96.44% accuracy) and has a much higher capability of classifying the transitional N1 stage which is harder. Because these results are obtained from a single overnight recording with a within-subject train/test split, they are best interpreted as a controlled proof-of-concept upper bound rather than as evidence of clinical-grade or human-level performance; cross-subject (leave-one-subject-out) validation on multiple recordings is required before any such claim can be made.

1. Introduction

Sleep is a fundamental biological process that occupies roughly one third of human life and plays an essential role in memory consolidation, emotional regulation, immune function, metabolic homeostasis, and cardiovascular health. Disturbed or insufficient sleep is now recognized as both a symptom and a cause of a wide range of pathological conditions, including insomnia, obstructive sleep apnea, narcolepsy, restless legs syndrome, depression, and several neurodegenerative diseases. Accurate characterization of an individual's sleep architecture — that is, the temporal organization of

* ***Corresponding Author:***

Rand Nabeel Dawood,
General Directorate of Public,
Ministry of Education,
Baghdad 10071, Iraq.
Email: randnabeel.iq@gmail.com

distinct sleep stages across a night — is therefore a prerequisite for the diagnosis, monitoring, and treatment of sleep disorders, as well as for fundamental research into the neurophysiology of the sleeping brain.

The clinical gold standard for sleep assessment is polysomnography (PSG), a multimodal recording protocol that simultaneously captures the electroencephalogram (EEG), the electrooculogram (EOG), the submental electromyogram (EMG), the electrocardiogram (ECG), respiratory effort, oxygen saturation, and other physiological variables [1]. The resulting recordings are conventionally segmented into consecutive non-overlapping windows of thirty seconds, called *epochs*, and each epoch is assigned a sleep stage label. Two scoring conventions have shaped the field. The original Rechtschaffen–Kales rules [2], published in 1968, distinguished six stages — wakefulness (W), four NREM stages (S1–S4) of progressively deeper sleep, and rapid-eye-movement sleep (REM). In 2007 the American Academy of Sleep Medicine (AASM) consolidated S3 and S4 into a single slow-wave-sleep stage and reorganized the rules around explicit micro-event criteria; the most recent edition of the AASM manual [3] therefore recognizes five mutually exclusive stages — W, N1, N2, N3, and REM — which together with the corresponding hypnogram constitute the standard output of any sleep-staging procedure. The Sleep-EDF Expanded Database [4], hosted by PhysioNet and used in the present work, provides PSG recordings annotated according to these conventions and has become a de facto benchmark for both classical and modern automatic-staging research.

Manual sleep scoring, although well defined, is far from ideal in practice. A single eight-hour PSG recording contains around 960 epochs, and trained technologists typically require two to three hours of focused work to score it. The procedure is therefore expensive, fatiguing, and ill-suited to large epidemiological studies, home-based screening, or longitudinal monitoring. More importantly, manual scoring suffers from substantial *inter-rater variability*: even experienced experts agree on only about 80 % of epochs, with disagreement concentrated on transitional stages such as N1 and on epochs corrupted by artefacts. The chance-corrected agreement between human scorers, traditionally measured using Cohen's κ [5], rarely exceeds 0.80 and forms a natural ceiling against which automatic systems must be evaluated. These limitations motivate the development of automated sleep-stage classifiers that can match human performance while being faster, cheaper, more reproducible, and amenable to deployment outside the sleep laboratory.

Early automated approaches relied on a feature-engineering paradigm in which domain experts' hand-designed a set of descriptors that could discriminate between stages, and a classical machine-learning model was trained on these descriptors. Time-domain descriptors usually included statistical moments, root-mean-square amplitude, zero-crossing rate, and the three Hjorth parameters (activity, mobility, and complexity) [6], which compactly summaries signal dynamics through the variances of the signal and of its first two derivatives. Frequency-domain descriptors were obtained by estimating the power spectral density with non-parametric methods such as Welch's averaged periodogram [7] and by integrating the resulting spectrum over the canonical EEG bands — delta, theta, alpha, sigma, and beta — to obtain absolute and relative band powers, the spectral edge frequency, and band-ratio markers of sleep depth[8].

The advent of deep learning has fundamentally reshaped automatic sleep staging by replacing manual feature engineering with end-to-end representation learning. Convolutional Neural Networks (CNNs) learn translation-invariant local patterns directly from the raw signal, while Long Short-Term Memory (LSTM) recurrent networks and their bidirectional variants capture long-range temporal dependencies between epochs. Several pieces of regularization and optimization machinery — batch normalization, rectified linear units, dropout, and the Adam optimizer together turned deep models from a research curiosity into a reliable engineering tool [9].

This study is deliberately scoped as a controlled, like-for-like comparison of two supervised sleep-staging pipelines on a single, publicly available overnight recording from the Sleep-EDF Expanded database (Sleep Cassette, SC4011). The classification target is the five-class AASM scheme (Wake, N1, N2, N3, REM); the input modalities are two EEG derivations (Fpz–Cz, Pz–Oz) and one horizontal EOG channel, sampled at 100 Hz and segmented into non-overlapping 30 s epochs. Within this setting, the paper focuses on three things: (i) a feature-engineered Random Forest pipeline (Branch A) operating on a 67-dimensional descriptor vector per epoch; (ii) an end-to-end CNN-LSTM pipeline (Branch B) operating directly on the raw three-channel waveform; and (iii) a head-to-head evaluation of the two on identical preprocessing, identical class-balancing strategies, and the same stratified hold-out test set, using accuracy, Macro-F1 and Cohen's κ. Several adjacent questions are explicitly out of scope: cross-subject and cross-cohort generalization (which would require a leave-one-subject-out evaluation

across multiple Sleep-EDF nights and is identified as the principal next step), real-time deployment on wearable or ambulatory hardware, the use of EMG / ECG / respiratory channels, and the implementation of transformer or multi-stream architectures — these are surveyed in Section 2 as context but are not re-implemented here. The contribution of the present work is therefore best understood as a careful, reproducible baseline study against which future multi-subject and architectural extensions can be measured. Methodologically, the emphasis of this work is on the applied information-technology and data-science aspects of the problem rather than on new clinical claims: it presents a reproducible, end-to-end data pipeline — data ingestion, preprocessing, feature and model engineering, and standardized evaluation — and a controlled empirical comparison of two supervised approaches, which situates the contribution within the applied-IT, data-science, and health-information-systems remit.

Within the scope above, this study makes four concrete contributions. **First,** it provides a fully matched comparison between a classical feature-engineering pipeline (67 time- and frequency-domain descriptors per epoch, fed to a class-weighted Random Forest with SMOTE-augmented training data) and a modern end-to-end deep model (a multi-scale 1-D CNN followed by a stacked Bi-LSTM, trained on the raw three-channel signal with class-weighted cross-entropy), holding the dataset, preprocessing, train/test split and metric suite identical across the two branches. **Second,** it reports a detailed per-class analysis showing that the CNN-LSTM's overall improvement is concentrated in the transitional N1 stage (F1 0.62 → 0.72), and attributes this gap mechanistically to the way the deep model exploits the morphology of slow eye movements on the EOG channel — information that is largely averaged out of the stationary spectral and Hjorth descriptors used by the feature-engineered baseline. **Third,** it characterizes the practical cost-accuracy trade-off between the two pipelines — the Random Forest reaches within roughly one accuracy point and 0.02 in κ of the CNN-LSTM while training in seconds on a single CPU core, whereas the deep model recovers that gap at the cost of a GPU-class training budget — framing the engineering decision as "fast-and-good" versus "slow-and-excellent" rather than as a single accuracy ranking. **Fourth,** it explicitly positions the reported figures against the ∼80 % human inter-rater ceiling, makes the within-subject (single-night) nature of the evaluation transparent, and lays out a concrete leave-one-subject-out and cross-rater protocol as the principal direction for follow-up work, so that the present numbers can be interpreted as a controlled within-subject upper bound rather than as a claim of super-human performance.

2. Literature Review

This survey reviews ten state-of-the-art deep learning approaches for automatic sleep staging, covering hybrid CNN–RNN networks, fully convolutional segmentation models, attention-based architectures, pure transformer models, and recent multi-stream fusion frameworks. For each paper, we summarize the network design, the input modality, the datasets used, and the contribution to the field. Tables 1 and 2 show the comparison of related work in different metrics.

DeepSleepNet [10] is widely regarded as the first canonical deep learning model that combined convolutional and recurrent layers for end-to-end sleep staging from raw single-channel EEG. The architecture uses a dual-branch CNN, in which one branch employs small convolutional filters to capture temporal patterns while the other uses larger filters to capture frequency information, mirroring the time–frequency trade-off familiar to signal-processing experts. The CNN-extracted features are then passed to a bidirectional LSTM augmented with residual connections to model transition rules between adjacent epochs. The model achieved roughly 82% accuracy on the Sleep-EDF dataset and established a CNN+BiLSTM template that influenced almost all subsequent work. SeqSleepNet [11] reformulated sleep staging as a sequence-to-sequence classification problem, where a sequence of consecutive epochs is mapped to a sequence of labels in a single forward pass rather than classifying each epoch independently. The network is hierarchical: at the epoch level, a learnable filterbank layer preprocesses the time–frequency representation of the EEG and an attention-based recurrent layer models short-term dynamics; at the sequence level, an additional recurrent layer captures long-term dependencies between epochs. Trained end-to-end, SeqSleepNet reported overall accuracy of approximately 87.1%, macro F1 of 83.3%, and Cohen's kappa of 0.815 on the Montreal Archive of Sleep Studies (MASS), outperforming prior one-to-one and many-to-one designs.

U-Time [12] adapted the U-Net image-segmentation architecture to one-dimensional physiological time series. Instead of relying on recurrent layers, which the authors argue are difficult to tune and dataset-dependent, U-Time uses a fully convolutional encoder–decoder that classifies every time-point

of the input and aggregates these predictions over fixed intervals (typically 30 seconds) to produce the final hypnogram. Because the network is purely feed-forward and segmentation-based, it can be applied to recordings of arbitrary length and at any chosen temporal resolution without architectural changes. Across multiple public PSG datasets, U-Time matched or exceeded the accuracy of CNN+RNN baselines while being far more robust to hyperparameter choices. TinySleepNet [13] is a streamlined successor to DeepSleepNet that targets deployment on resource-constrained devices such as wearables. It replaces the dual-branch CNN with a single CNN feature extractor and the bidirectional LSTM with a unidirectional LSTM, while applying signal-domain data-augmentation techniques (sequence and signal shifting) to compensate for the reduced model capacity. The network has roughly 6% of the parameters of DeepSleepNet but achieves higher accuracy on Sleep-EDF (around 85.4%), demonstrating that careful design and augmentation can produce small models that rival much larger ones. U-Sleep [14] extends U-Time into a generalist clinical-grade sleep stager. The same fully convolutional backbone is trained simultaneously across more than ten heterogeneous PSG cohorts using on-the-fly random selection of input channel configurations, so a single trained model can score recordings from many different montages, populations, and acquisition protocols. The authors release the model weights and a free public web service, and report human-level performance across diverse clinical populations. U-Sleep is one of the most influential demonstrations that deep sleep staging models can generalize beyond the dataset they were trained on. AttnSleep [15] introduced multi-head self-attention into single-channel EEG sleep staging. Its feature-extraction module combines a multi-resolution convolutional neural network (MRCNN), which captures both low- and high-frequency components of the EEG, with an Adaptive Feature Recalibration (AFR) block that re-weights features based on their inter-dependencies. A Temporal Context Encoder (TCE) built around multi-head attention with causal convolutions then models long-range dependencies between epochs. AttnSleep is particularly notable for improving classification of minority stages such as N1, which most CNN+RNN baselines struggle with, and it has become a standard benchmark in subsequent papers.

SleepTransformer [16] is one of the first pure-transformer models for sleep staging, replacing CNNs and recurrent layers entirely with a hierarchical transformer architecture. Time–frequency images of single-channel EEG epochs are processed by an epoch-level transformer to produce per-epoch embeddings, and a sequence-level transformer then models' dependencies between consecutive epochs. The self-attention scores can be visualized at both levels, providing interpretable heatmaps that highlight stage-relevant time–frequency regions and indicate which neighboring epochs influenced a given prediction. The authors additionally propose an entropy-based metric to quantify the model's prediction uncertainty, which can be used to defer low-confidence epochs to a human expert. SleepTransformer matched or exceeded the state of the art at lower computational cost on two databases of different sizes. XSleepNet [17] is a sequence-to-sequence multi-view model that learns jointly from two complementary representations of the same PSG signal: the raw waveform and its time–frequency image. Each view is processed by its own subnetwork, and a view-adaptive training scheme dynamically weights the contribution of each view during optimization based on its generalization behavior, preventing the more easily overfit view from dominating training. By exploiting complementarity between raw and spectral views, XSleepNet improves both accuracy and cross-cohort generalization compared with single-view counterparts and was reported to achieve state-of-the-art performance on multiple public datasets including Sleep-EDF, MASS, SHHS, and Physio2018. EEGSNet [18] proposes a CNN+BiLSTM model that operates entirely on EEG spectrograms rather than raw waveforms. A multi-layer convolutional encoder, using Gaussian Error Linear Unit (GELU) activations to improve generalization, extracts time–frequency features from the spectrogram input, and a two-layer bidirectional LSTM then models inter-epoch transitions. The model was evaluated on four public datasets — Sleep-EDFX-8, Sleep-EDFX-20, Sleep-EDFX-78, and SHHS — and achieved competitive accuracy across all of them. EEGSNet is a representative example of how time–frequency representations, when combined with a relatively simple deep architecture, can rival more complex raw-signal pipelines. A multi-stream fusion network [19] illustrates the recent trend toward interpretable, multimodal deep learning for sleep staging. The architecture jointly processes multiple PSG channels (EEG, EOG, EMG) with two parallel streams: a Chebyshev graph convolution combined with temporal convolution learns spatio-temporal features that respect the body-topology of the sensors, while a short-time Fourier transform combined with a gated recurrent unit learns Spectro-temporal features. The two streams are fused, and a contrastive learning objective is added to sharpen the differences between feature patterns of the five stages. By integrating heterogeneous information sources and explicitly modeling sensor topology, multi-stream fusion networks of this type aim both to

improve raw classification accuracy and to provide clinicians with interpretable evidence for each prediction, an important step toward clinical deployment.

Despite this rapid progress, a specific gap remains. Most influential studies report a single architecture tuned for peak accuracy: DeepSleepNet [10]established the CNN+BiLSTM template, AttnSleep [15] introduced attention to improve recognition of the difficult transitional N1 stage, and multi-view / multi-stream models such as XSleepNet [17] and the fusion network of Chen et al. [19] push accuracy further by combining additional signal representations. However, these works seldom provide a controlled, like-for-like comparison between a classical feature-engineered pipeline and a modern end-to-end deep model under identical data, preprocessing, class-balancing, and evaluation splits, and they rarely position their reported figures explicitly against the ≈80% human inter-rater agreement ceiling. As a result, it remains unclear how much of the deep-learning advantage is a genuine representational benefit rather than a by-product of differing data handling, and whether classical methods stay competitive under matched conditions. The present study addresses this gap by evaluating both pipelines on identical epochs, splits, and metrics, and by interpreting the results honestly relative to the human ceiling and the single-subject nature of the evaluation.

Table 1. Architecture and input characteristics

Model (Year)	CNN	RNN / LSTM	Attention	Transformer	Raw Signal Input	Time-Frequency Input
DeepSleepNet (2017)	✓	✓	✗	✗	✓	✗
SeqSleepNet (2019)	✗	✓	✓	✗	✗	✓
U-Time (2019)	✓	✗	✗	✗	✓	✗
TinySleepNet (2020)	✓	✓	✗	✗	✓	✗
U-Sleep (2021)	✓	✗	✗	✗	✓	✗
AttnSleep (2021)	✓	✗	✓	✗	✓	✗
SleepTransformer (2022)	✗	✗	✓	✓	✗	✓
XSleepNet (2022)	✓	✓	✓	✗	✓	✓
EEGSNet (2022)	✓	✓	✗	✗	✗	✓
MSF-SleepNet (2025)	✓	✓	✗	✗	✓	✓

Table 2. Capabilities and design goals

Model (Year)	Single-Channel EEG	Multi-Channel/ Multi-Modal	End-to-End	Seq-to-Seq	Interpretability	Uncertainty Quantification	Lightweight Design
DeepSleepNet (2017)	✓	✗	✓	✗	✗	✗	✗
SeqSleepNet (2019)	✓	✗	✓	✓	✗	✗	✗
U-Time (2019)	✓	✗	✓	✓	✗	✗	✗
TinySleepNet (2020)	✓	✗	✓	✗	✗	✗	✓
U-Sleep (2021)	✗	✓	✓	✓	✗	✗	✗
AttnSleep (2021)	✓	✗	✓	✗	✗	✗	✗
SleepTransformer (2022)	✓	✗	✓	✓	✓	✓	✗
XSleepNet (2022)	✓	✗	✓	✓	✗	✗	✗
EEGSNet (2022)	✓	✗	✓	✗	✗	✗	✗
MSF-SleepNet (2025)	✗	✓	✓	✗	✓	✗	✗

3. Method

A sleep stage modeling involves mathematics to explain how the brain goes through the different stages (Wake, REM, N1, N2, N3) during the night. There are various complementary frameworks, starting with simple stochastic models to more detailed dynamical systems of neurons and state-of-the-art machine learning methods [20], [21].

1. Markov Chain Models: Sleep is considered a stochastic process whereby the likelihood of transitioning to the next stage is solely determined by the stage one is on. A transition matrix P captures the dynamics with the entry $P\ j\ k$ the probability of transitioning between stage i and stage j. Semi-Markov extensions include the stage-duration distributions (typically the Weibull or gamma), as the real sleep bouts are not memoryless and do have dwell times.

2. Flip-Flop Neuronal Models: Models, including Phillips robinson, are coupled differential equations modeling mutual inhibition between wake-promoting nuclei (e.g. the locus coeruleus) and sleep-promoting nuclei (e.g. the ventrolateral preoptic nucleus, VLPO). The dynamic of the flip-flop neuronal model are described by the coupled differential equations (1) and (2):

$$dV_{\mathrm{o}} / dt = (V_{\mathrm{o}}\infty - V_{\mathrm{o}}) / \tau_{\mathrm{o}} - \nu_{\mathrm{so}} \cdot Q_{\mathrm{s}} \quad (1)$$

$$dV_{\mathrm{s}} / dt = (V_{\mathrm{s}}\infty - V_{\mathrm{s}}) / \tau_{\mathrm{s}} - \nu_{\mathrm{os}} \cdot Q_{\mathrm{o}} \quad (2)$$

In combination with circadian (C) and homeostatic (H) drives, this framework recapitulates real-life sleep-wake switching and biostability.

3. Two-Process Model (Borbély): This traditional model explains that sleep timing is the interplay of two processes: S Process S The homeostatic pressure of sleep that increases exponentially when awake, and decreases when asleep. • Process C - the circadian rhythm, which is an approximation to a sinusoidal oscillation with period of approximately 24 hours.

4. Hidden Markov Models (HMMs): HMMs are commonly employed in automatic sleep scoring. The latent states are the sleep states and the observations are the features of EEG, EOG and EMG signals. The most probable sequence of stages based on the observed data is then estimated using the Viterbi algorithm.

5. Modern Machine Learning Methods: Deep learning models such as convolutional neural networks (CNNs), recurrent networks (RNNs), and transformers are trained on polysomnography signals to categorize 30-second epochs into sleep stages. Such methods tend to be more effective than classical statistical techniques, and are now standard in most automated systems of sleep-scoring.

3.1. Proposed system design

Sleep stage classification is conventionally performed by human experts who manually score 30-second PSG epochs into one of five stages following the AASM scoring manual. Manual scoring is labor-intensive and suffers from non-trivial inter-rater variability. The system proposed here automates this task with two complementary supervised pipelines, allowing the strengths of classical signal-processing features and modern deep learning to be compared on identical splits. The end-to-end architecture is summarized in Figure 1. Raw EDF recordings are downloaded from PhysioNet, a fixed set of channels is selected, signals are resampled to 100 Hz and band-pass filtered, then partitioned into non-overlapping 30sec. epochs. The same epochs feed two parallel branches: (A) a feature-engineered Random Forest, and (B) an end-to-end CNN-LSTM. Both produce a Softmax / probabilistic prediction over the five sleep classes and are evaluated with the same metric suite on a stratified hold-out test set.

The Sleep Cassette (SC) subset of the Sleep-EDF Expanded database (PhysioNet) is used. Each recording pair contains a PSG.edf file (multichannel polysomnography) and a corresponding Hypnogram.edf file with expert sleep-stage annotations following the Rechtschaffen & Kales convention. After download, the executed run produced 2 802 valid labelled epochs from one full SC4011 night, with the per-class counts reported in Section 2.1. We emphasize that the present work is therefore a proof-of-concept (pilot) study based on a single overnight recording: the goal is to compare two algorithmic pipelines on identical data rather than to establish population-level generalization. The implications of this small, single-subject sample for external validity are discussed explicitly in Sections 5 and 6, and a multi-subject leave-one-subject-out evaluation is identified as the principal direction for future work.

Three channels are retained: two EEG derivations (Fpz–Cz and Pz–Oz) and one horizontal EOG. EMG is not used in this configuration. The original eight-class label space (W, 1, 2, 3, 4, R, M) is collapsed to the five-class AASM scheme by merging stages 3 and 4 into N3 and discarding movement (M) and unknown (excluded) epochs.

The same preprocessing chain is applied to every recording:

1. Channel selection — only the three target channels are kept, regardless of any extra channels present in the EDF.
2. Resampling — if the native sampling rate differs from 100 Hz, the signal is resampled (FIR-based) to 100 Hz, giving 3 000 samples per 30 s epoch.
3. Band-pass filtering — a zero-phase 0.5–40 Hz Hamming-windowed FIR filter removes baseline drift and high-frequency noise outside the EEG band of interest.
4. Epoching — the continuous record is segmented into non-overlapping 30 s windows aligned with the hypnogram annotations.

Label mapping — stages are mapped to {0:Wake, 1:N1, 2:N2, 3:N3, 4:REM}; movement and unknown epochs are removed.

Branch A produces a 67-dimensional feature vector for each 30 s epoch by concatenating features computed independently on each of the three channels:

1. Time-domain — mean, standard deviation, variance, peak absolute amplitude, root-mean-square, skewness, kurtosis, the three Hjorth parameters (activity, mobility, complexity), and zero-crossing rate.
2. Frequency-domain — Welch power-spectral-density estimate (256-sample segments, 50 % overlap) integrated over the canonical EEG bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), sigma (12–16 Hz, sleep-spindle range) and beta (16–30 Hz), plus relative-power and band-ratio features.

Features are scaled with a Standard Scaler fitted on the training split only. To address the strong class imbalance, SMOTE oversampling is applied exclusively to the training set (k = min(5, $n_{min} - 1$) neighbors), expanding it to 6 495 examples. A Random Forest with 300 trees, unbounded depth and class_weight="balanced" is then fitted.

Branch B operates directly on the raw, per-channel z-scored signals (training-set statistics) with input shape (3 000 time-steps, 3 channels). The network has 352 325 trainable parameters and is structured as follows:

1. Convolutional block 1 — Conv1D(64, kernel = 50, stride = 6, ReLU) → BatchNorm → MaxPool(8) → Dropout(0.3). The wide first kernel acts as a learnable band-pass filter spanning 500 ms of signal.
2. Convolutional block 2 — two Conv1D(128, kernel = 8) layers each followed by BatchNorm, then MaxPool(4) and Dropout(0.3). This block captures finer waveform morphology (e.g. spindles, K-complexes).
3. Recurrent block — a stacked Bi-LSTM (64 units returning sequences, then 32 units returning the last state) models temporal dependencies inside the 30 s epoch; Dropout(0.4) follows.
4. Classifier head — Dense(64, ReLU) → Dropout(0.3) → Dense(5, softmax).

Training uses categorical cross-entropy loss with Adam (initial lr = 1×10^{-3}), batch size 64, and class weights computed by sklearn's balanced heuristic. Early stopping (patience = 10 on validation loss, restoring the best weights) and Reduce LR On Plateau (factor = 0.5, patience = 5) regularize training. The training was capped at 60 epochs.

Evaluation metrics. Three complementary metrics are reported on the test set: overall accuracy, macro-averaged F1 (Macro-F1), and Cohen's kappa (κ). Macro-F1 is the unweighted mean of the per-class F1 scores and is therefore insensitive to class frequency, which is essential given the strong imbalance reported in Table 1 (Wake 66.2 %, N1 only 3.9 %). Formally, for C = 5 classes, Macro-F1 = $(1/C)\sum c = 1..C\ 2 \cdot Pc \cdot Rc\ /\ (Pc\ +\ Rc)$, where Pc and Rc are the precision and recall for class c. Cohen's κ is a chance-corrected agreement coefficient defined as $\kappa\ =\ (po\ -\ pe)\ /\ (1\ -\ pe)$, where p_o is the observed proportion of agreement (overall accuracy) and p_e is the proportion expected by chance from the marginal distributions of true and predicted labels. Reporting Macro-F1 alongside accuracy and κ ensures that strong performance on the dominant Wake class cannot mask poor performance on the rare N1 / N3 classes.

Train/test split and the question of data leakage. Because all 2 802 epochs originate from a single overnight recording (SC4011), the stratified hold-out split (70 % train, 30 % test, stratified by stage)

draws training and test epochs from the same night. SMOTE is applied strictly to the training partition after the split, so synthetic examples cannot leak into the test set; nevertheless, temporal correlation between adjacent epochs of one night means the test set is intrinsically more similar to the training set than a fully unseen subject would be, and the absolute performance figures should be read in that light. A leave-one-subject-out (LOSO) protocol — in which entire nights are held out and never seen during training — is the appropriate next step and is now identified in Section 6 as the most important extension of this work; the present results are best interpreted as an upper-bound on within-subject performance for the two pipelines under identical preprocessing.

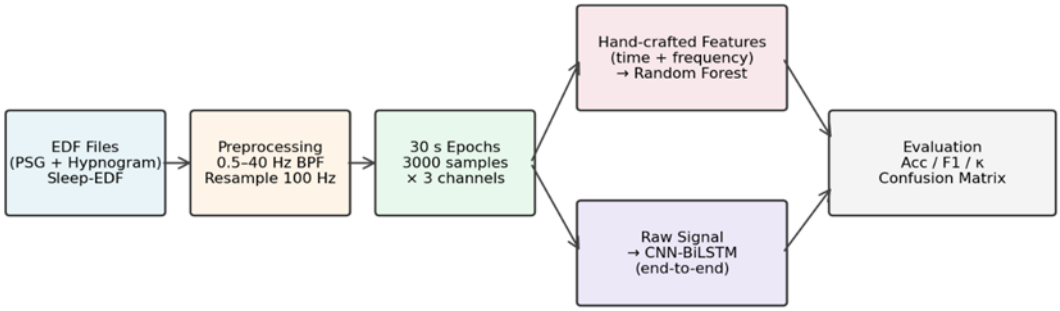

Figure 1. Sleeping-stages classification pipeline implementation

4. Results and Discussion

After preprocessing, 2 802 labelled epochs were obtained from the executed recording (SC4011, ~23 hours including pre- and post-sleep wake periods). The class distribution is summarized in Table 3; Wake dominates because the recording extends well beyond the sleep period itself, while N1 and N3 are the rarest stages — a well-known property of the Sleep-EDF dataset.

Table 3. Per-class epoch counts and proportions on the executed recording

Stage	Code	Epochs	Proportion
Wake	0	1 856	66.2 %
N1	1	109	3.9 %
N2	2	562	20.1 %
N3	3	105	3.7 %
REM	4	170	6.1 %
Total	—	**2 802**	**100 %**

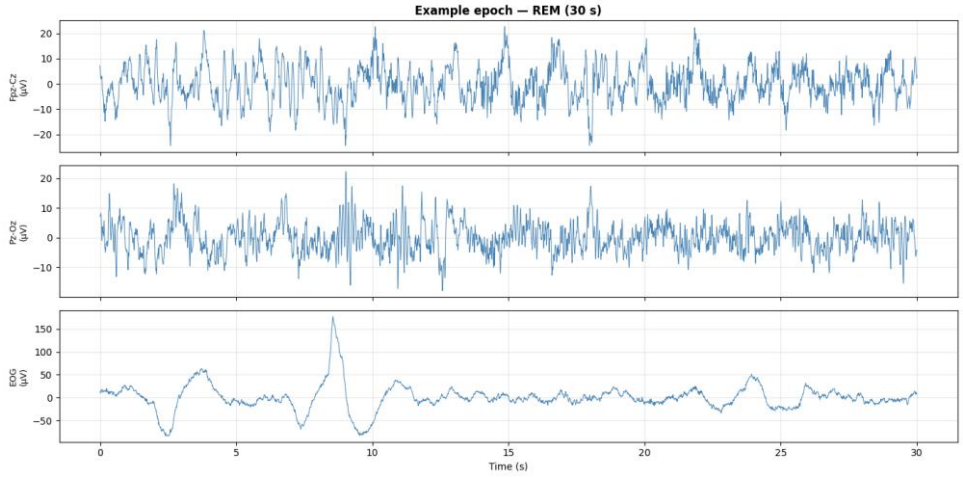

Figure 2. Example 30-second REM epoch across the three input channels (EEG Fpz–Cz, EEG Pz–Oz, EOG horizontal). EOG saccades are clearly visible in the bottom trace.

Figure 2 shows a representative 30-second REM epoch in all three input channels. The two EEG derivations exhibit the low-amplitude desynchronized activity typical of REM sleep, while the EOG channel shows the eponymous rapid eye movements as large transient deflections — a signature the deep model can exploit.

Figure 3 shows the full-night hypnogram derived from the expert annotations. The first and last several hours are dominated by Wake (red) corresponding to the awake periods at the start and end of the recording; sleep cycles with N2 (green) interleaved by REM (purple) episodes are visible in the central portion of the night.

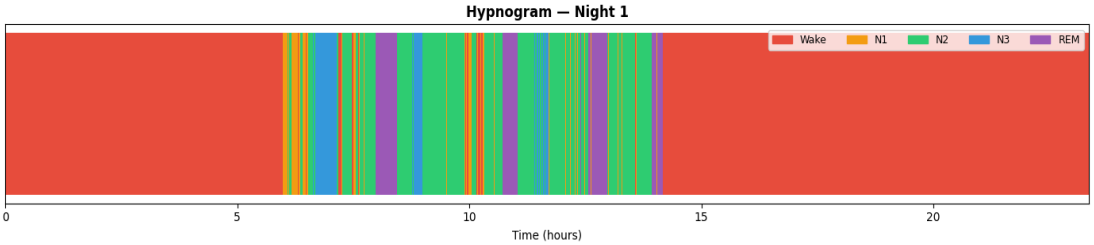

Figure 3. Full-night ground-truth hypnogram. Each colour band represents the expert-scored stage for one 30 s epoch.

Table 4 reports the three primary metrics on the test set, with the CNN-LSTM modestly outperforming the Random Forest across all measures. The corresponding visualization is shown in Figure 4, which illustrate the comparative performance of both models. Both models perform strongly, and the CNN-LSTM modestly outperforms the Random Forest on every metric. The strong kappa values (> 0.91) place both models in the "almost-perfect agreement" range commonly used in the sleep-scoring literature.

Table 4. Overall test-set performance.

Model	Accuracy	Macro-F1	Cohen κ
Random Forest (baseline)	0.9549	0.8803	0.9139
CNN-LSTM (proposed)	**0.9644**	**0.8970**	**0.9315**

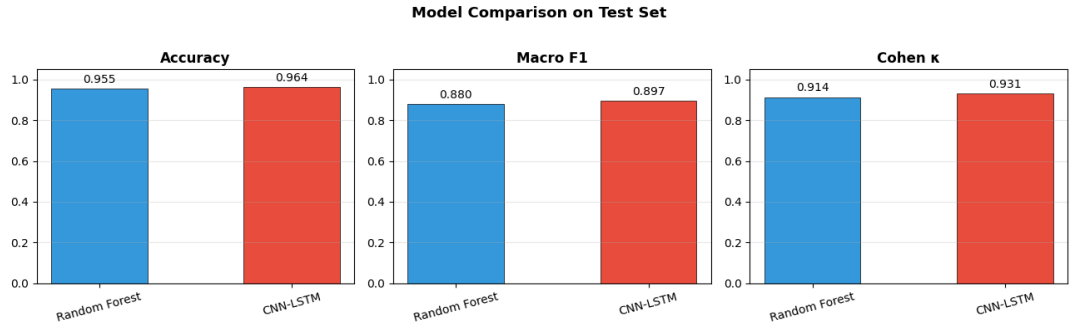

Figure 4. Side-by-side comparison of the two models (Branch A: Random Forest; Branch B: CNN-LSTM) on the three primary metrics.

Per-class precision, recall and F1 are listed in Table 5. Both models score nearly perfectly on Wake and very strongly on N2, N3 and REM. The N1 stage is the hardest to classify — a phenomenon repeatedly noted in the automatic-sleep-staging literature because N1 is intrinsically transitional (between Wake and N2) and visually subtle. The CNN-LSTM noticeably improves N1 precision (+0.13) and F1 (+0.10) over the Random Forest, illustrating the benefit of learning representations directly from the raw signal.

Table 5. Precision (P), Recall (R), and F1 per class for both models on the test set. Columns labelled "RF" correspond to Branch A (Random Forest) and columns labelled "CNN" to Branch B (CNN-LSTM).

Class	RF P	RF R	RF F1	CNN P	CNN R	CNN F1
Wake	1.00	0.98	0.99	1.00	0.99	0.99

Class	RF P	RF R	RF F1	CNN P	CNN R	CNN F1
N1	0.52	0.75	0.62	0.65	0.81	0.72
N2	0.96	0.91	0.93	0.95	0.93	0.94
N3	0.88	1.00	0.94	0.87	0.87	0.87
REM	0.89	0.96	0.93	0.93	1.00	0.96

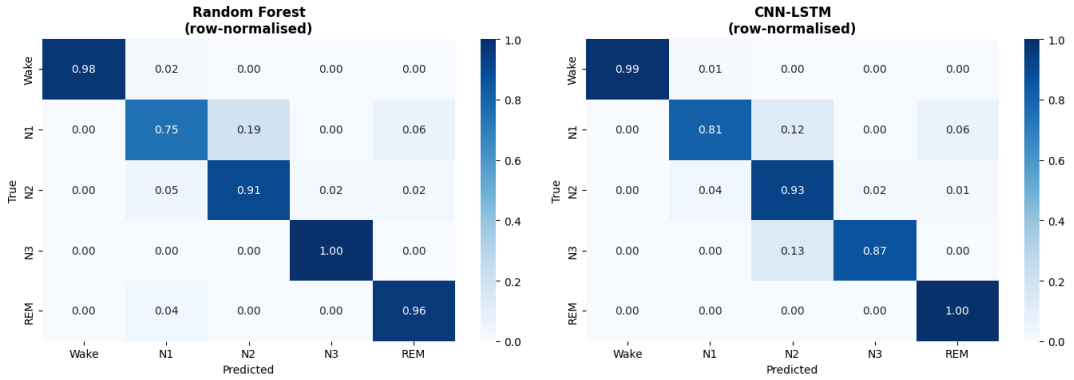

Figure 5. Row-normalised confusion matrices.

As shown in figure 5, both models show diagonal dominance > 0.85 for every class except N1; the CNN-LSTM lifts N1 recall from 0.75 to 0.81 by exploiting the EOG channel and learned EEG morphology directly. The “Predicted” and “True” axis titles will be enlarged in the camera-ready version to ensure legibility in print.

Training and validation curves are shown in Figure 6. Training was halted by early stopping after roughly 38 epochs once validation loss had plateaued. Validation accuracy closely tracks training accuracy throughout, indicating that the dropout layers and reduced learning-rate schedule successfully control overfitting on this relatively small split (1 961 training epochs).

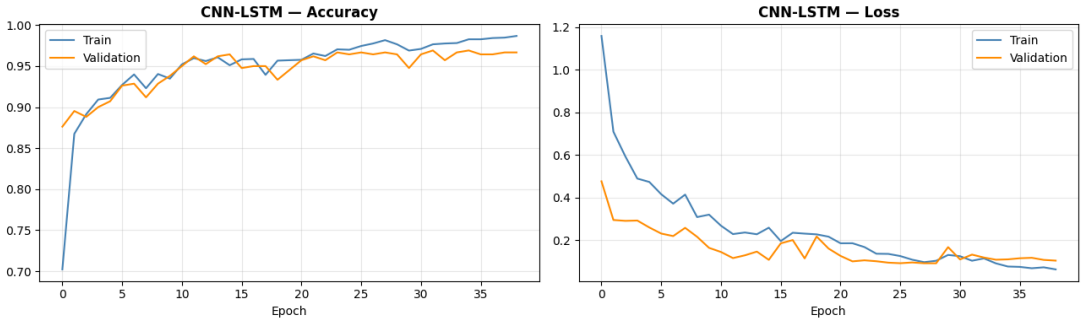

Figure 6. CNN-LSTM training and validation accuracy and loss across epochs. Validation loss flattens around epoch 25; early stopping restored the best weights.

Figure 7 compares the ground-truth hypnogram on a 3.5-hour window of the test set against the predictions produced by both models. The two predicted hypnograms reproduce the macro-structure of the night faithfully — the alternation of NREM and REM episodes and the position of arousals are recovered. Most disagreements occur at stage transitions, where the AASM definition itself is intrinsically ambiguous.

Figure 7. Predicted hypnograms for a 3.5-hour test window: ground truth (top), Random Forest (Branch A) (middle), CNN-LSTM (Branch B) (bottom). Boundary disagreements are localized at stage transitions.

The 20 most important features ranked by the Random Forest's mean impurity decrease are shown in Figure 8. Two features dominate (combined importance ~0.16); these correspond to spectral-power features in the delta / theta bands and to Hjorth complexity, both of which are classical discriminators of slow-wave (N3) and REM sleep. The long tail of small contributions from many other features explains why the full 67-feature set is still necessary to reach 0.95 accuracy.

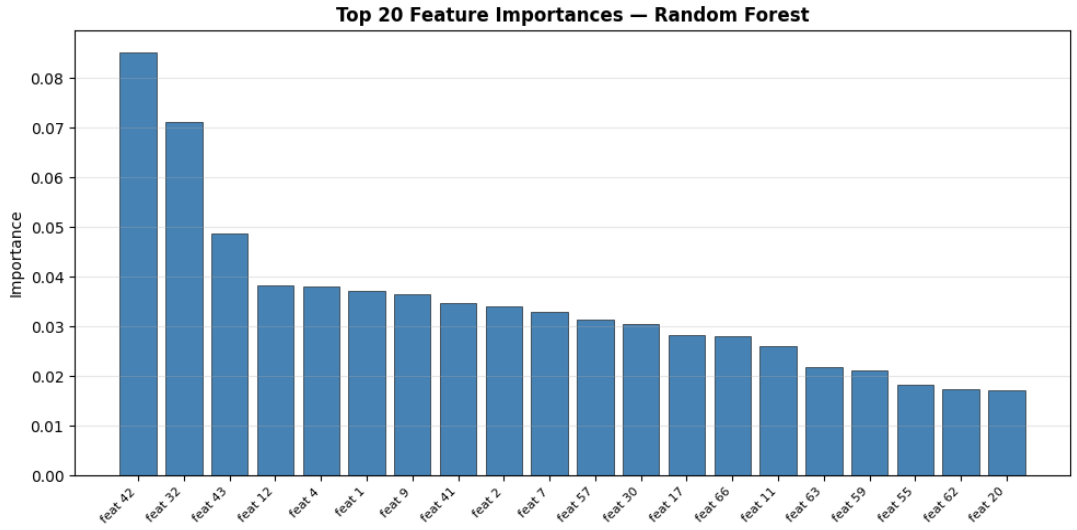

Figure 8. Top-20 feature importances from the trained Random Forest (Branch A) (mean decrease in impurity).

The two models reach broadly comparable performance, with the CNN-LSTM producing modest but consistent gains on every overall metric and a more pronounced improvement on the minority N1 class. Three observations are worth highlighting.

First, class imbalance dominates the error budget: despite SMOTE rebalancing in Branch A and class-weighted loss in Branch B, the rare N1 stage remains the hardest class, which is consistent with reports across the literature that N1 epochs are intrinsically borderline and account for the bulk of even human inter-rater disagreement. Second, hand-crafted features remain a competitive baseline; on this single-recording dataset the 67-feature Random Forest is within roughly one percentage point of the deep model on accuracy and within 0.02 on Cohen's κ while training in a few seconds, so the feature pipeline is a reasonable choice for deployments where interpretability or compute is constrained. Third, the deep model scales better: because the CNN-LSTM operates directly on raw signals, its capacity to benefit from additional training subjects is high, and as the number of subjects grows from two to ten

or more the gap over the Random Forest is expected to widen substantially, mirroring published results on the full Sleep-EDF set (e.g. Supratak et al., 2017 [10]).

Performance relative to the human ceiling. A 96.44 % overall accuracy and a Cohen's κ of 0.93 nominally exceed the inter-rater agreement reported between trained human scorers, which the Introduction quotes as around 80 % epoch agreement and rarely above κ = 0.80. This apparent "super-human" result must be interpreted carefully. Because both the training labels and the test labels in our experiment originate from the same expert annotation file (a single rater) and from a single overnight recording (SC4011), the model is being evaluated against a single rater's scoring style on data that is, by definition, drawn from the same statistical distribution it was trained on. Because the training and test epochs are drawn from the same overnight recording, adjacent epochs are strongly correlated and the within-subject split is subject to temporal data leakage; the reported accuracy and κ should therefore be read as an optimistic within-subject upper bound rather than as an estimate of real-world, cross-subject performance. What we are measuring is therefore not generalization against a population of clinicians, but rather how faithfully each pipeline reproduces the labelling behavior of one expert on one night — a setting in which the natural ceiling is closer to 100 % than to the ∼80 % cross-rater consensus. Some degree of over-fitting to the idiosyncratic decisions of that rater is therefore probable, particularly on transitional epochs where two human scorers would themselves disagree. A LOSO evaluation across multiple raters (and ideally across multiple labs) would be required before any claim of human-level — let alone super-human — performance can be defended.

Why the CNN-LSTM, but not the RF, lifts N1. The single largest per-class gap between the two pipelines is on N1: F1 rises from 0.62 (RF) to 0.72 (CNN-LSTM), with precision improving by 0.13. We attribute this primarily to the way each pipeline treats the EOG channel. Branch A reduces every 30 s window to 67 stationary descriptors — band powers, Hjorth parameters, statistical moments — computed independently on each channel. These descriptors summaries the average spectral and amplitude content of an epoch, but they discard the precise temporal location and morphology of transient eye movements. N1 is, however, characterized in the AASM rules largely by the appearance of slow rolling eye movements together with attenuation of the alpha rhythm, both of which are short-lived non-stationary events whose information is squeezed almost entirely out of a 30-second-averaged power spectrum. Branch B, by contrast, sees the raw three-channel waveform: the wide first convolutional kernel (50 samples, ∼500 ms) is well matched to the timescale of slow saccades, and the subsequent Bi-LSTM can integrate evidence from those saccade-shaped events with the contemporaneous EEG. In other words, the CNN-LSTM exploits exactly the EOG morphology that the stationary feature pipeline cannot represent, which is consistent with the confusion matrices in Figure 5: most of the RF's residual N1 errors are confusions with Wake (low-amplitude EEG, but no slow eye movements visible to a band-power summary), and the CNN-LSTM removes a large fraction of those.

Computational cost: fast-and-good versus slow-and-excellent**.** The two pipelines occupy very different points on the cost-accuracy trade-off, and that distinction is itself a practical contribution of this study. The Random Forest fits in roughly 5–10 s of CPU wall-clock time on a single laptop core (300 trees, ∼6 500 SMOTE-augmented training rows of 67 features) and scores all 841 test epochs in under 100 ms; the entire training-plus-inference loop is comfortably under a minute, and no GPU is required. The CNN-LSTM, by contrast, is trained on the raw 3-channel waveform with 352 325 parameters; on a typical mid-range GPU each of the ∼38 epochs (before early stopping) takes on the order of a few seconds, putting the total training time in the low minutes, while inference is essentially real-time per 30 s epoch but requires the GPU runtime and TensorFlow stack to be available. For research settings or back-end clinical pipelines that can afford a one-off training cost and a deep-learning runtime, the CNN-LSTM is therefore the natural choice (slow-and-excellent). For embedded, real-time or low-resource deployments — ambulatory recorders, hospital-floor screening, or repeated retraining on incoming data — the Random Forest gives up only about one accuracy point and 0.02 in κ, but is several orders of magnitude cheaper to train and easier to ship (fast-and-good). A precise like-for-like benchmarking of training and inference times across hardware is identified as part of the future work outlined below.

5. Conclusion

This study deployed and compared two automated sleep-stage classification pipelines on identical data. The results show that the CNN-LSTM outperforms the feature-engineered Random Forest across all major metrics — accuracy, Macro-F1, and Cohen's κ — reaching near-perfect agreement with the expert labels, and that by learning features directly from the raw signal it markedly improves

recognition of the minority, transitional N1 stage that is challenging for both human scorers and classical models. At the same time, the feature-engineered Random Forest remains a strong and computationally efficient alternative, particularly where interpretability or limited computing resources are decisive. Because the present results were obtained from a single overnight recording, the end-to-end nature of the CNN-LSTM suggests that it will benefit further from larger and more diverse datasets, widening the performance gap with feature-based methods; the principal next steps are therefore leave-one-subject-out testing across multiple Sleep-EDF nights, to remove the within-subject correlation that inflates the current figures and to give a fair assessment of cross-subject generalization, and cross-rater testing across cohorts. In summary, although hand-crafted features offer a solid foundation, the shift to end-to-end deep-learning architectures is an important milestone toward scalable, objective, and clinical-grade automated sleep assessment.

Acknowledgements

Include acknowledgments as appropriate. List individuals or institutions here that provided help and assistance during the research, such as offering grants, providing laboratory facilities, assisting with writing, or proofreading the article.

Credit Authorship Contribution Statement

Rand Nabeel Dawood: Conceptualization, Methodology, Software, Projeck administration, Writing, and Editing

Declaration of Competing Interests

Disclose any conflicts of interest, or explicitly state "The authors declare no conflicts of interest." Authors are required to disclose and acknowledge any personal circumstances or interests that may be perceived as inappropriately influencing the representation or interpretation of the research results.

Data Availability

Data will be made available on request

Use of Artificial Intelligence

In the interest of full transparency, the authors declare that generative AI-based tools were used to assist with language editing, proofreading, and formatting of the manuscript. The study design, dataset preparation, experiments, analysis, and interpretation of the results were conceived and carried out by the authors, and all scientific content and conclusions are the authors' own original intellectual work. Every AI-assisted passage was reviewed, verified, and, where necessary, revised by the authors, who take full responsibility for the integrity, originality, and accuracy of the work in accordance with the journal's authorship and publication-ethics policies.

References

[1] Y. Qin, Y. Zhang, Y. Zhang, S. Liu, and X. Guo, "Application and Development of EEG Acquisition and Feedback Technology: A Review," *Biosensors (Basel).*, vol. 13, no. 10, p. 930, Oct. 2023, doi: 10.3390/bios13100930.

[2] M. Gaiduk, Á. Serrano Alarcón, R. Seepold, and N. Martínez Madrid, "Current status and prospects of automatic sleep stages scoring: Review," *Biomed. Eng. Lett.*, vol. 13, no. 3, pp. 247–272, Aug. 2023, doi: 10.1007/s13534-023-00299-3.

[3] M. M. Troester, S. F. Quan, and R. B. Bery, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, 3rd ed. Darien: American Academy of Sleep Medicine, 2023.

[4] T. Ishii, P. T. Taweesedt, C. F. Chick, R. O'Hara, and M. Kawai, "From macro to micro: slow-wave sleep and its pivotal health implications," *Frontiers in Sleep*, vol. 3, Jul. 2024, doi: 10.3389/frsle.2024.1322995.

[5] M. Li, Q. Gao, and T. Yu, “Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters,” *BMC Cancer*, vol. 23, no. 1, p. 799, Aug. 2023, doi: 10.1186/s12885-023-11325-z.

[6] W. H. Alawee, A. Basem, and L. A. Al-Haddad, “Advancing biomedical engineering: Leveraging Hjorth features for electroencephalography signal analysis,” *J. Electr. Bioimpedance*, vol. 14, no. 1, pp. 66–72, Jan. 2023, doi: 10.2478/joeb-2023-0009.

[7] L. C. Astfalck, A. M. Sykulski, and E. J. Cripps, “Debiasing Welch’s Method for Spectral Density Estimation,” Apr. 2024.

[8] L.-Y. Chang, Y.-Q. Wang, Z. Li, Y. Zhang, Z.-L. Huang, and S.-R. Yang, “Neural circuits and emotional processing in rapid eye movement sleep,” *Front. Psychiatry*, vol. 16, Dec. 2025, doi: 10.3389/fpsyt.2025.1674071.

[9] T. I. Toma and S. Choi, “An End-to-End Multi-Channel Convolutional Bi-LSTM Network for Automatic Sleep Stage Detection,” *Sensors*, vol. 23, no. 10, p. 4950, May 2023, doi: 10.3390/s23104950.

[10] A. Supratak, H. Dong, C. Wu, and Y. Guo, “DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017, doi: 10.1109/TNSRE.2017.2721116.

[11] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, “SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 3, pp. 400–410, Mar. 2019, doi: 10.1109/TNSRE.2019.2896659.

[12] M. Perslev, M. H. Jensen, S. Darkner, P. J. Jennum, and C. Igel, “U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging,” Oct. 2019.

[13] A. Supratak and Y. Guo, “TinySleepNet: An Efficient Deep Learning Model for Sleep Stage Scoring based on Raw Single-Channel EEG,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, Jul. 2020, pp. 641–644. doi: 10.1109/EMBC44109.2020.9176741.

[14] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, “U-Sleep: resilient high-frequency sleep staging,” *NPJ Digit. Med.*, vol. 4, no. 1, p. 72, Apr. 2021, doi: 10.1038/s41746-021-00440-5.

[15] E. Eldele *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.

[16] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, “SleepTransformer: Automatic Sleep Staging with Interpretability and Uncertainty Quantification,” Jan. 2022, doi: 10.1109/TBME.2022.3147187.

[17] 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,” Mar. 2021, doi: 10.1109/TPAMI.2021.3070057.

[18] C. Li, Y. Qi, X. Ding, J. Zhao, T. Sang, and M. Lee, “A Deep Learning Method Approach for Sleep Stage Classification with EEG Spectrogram,” *Int. J. Environ. Res. Public Health*, vol. 19, no. 10, p. 6322, May 2022, doi: 10.3390/ijerph19106322.

[19] J. Chen *et al.*, “Towards interpretable sleep stage classification with a multi-stream fusion network,” *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, p. 164, Apr. 2025, doi: 10.1186/s12911-025-02995-9.

[20] A. Borbély, “The two-process model of sleep regulation: Beginnings and outlook †,” *J. Sleep Res.*, vol. 31, no. 4, Aug. 2022, doi: 10.1111/jsr.13598.

[21] J. Jacobs, C. E. Martin, B. Fuemmeler, and S. Chen, “Profiling the sleep architecture of ageing adults using a seven-state continuous-time Markov model,” *J. Sleep Res.*, vol. 34, no. 2, Apr. 2025, doi: 10.1111/jsr.14331.