

An Automated Disease Prediction System for Sleep Behaviour Analytics Using Deep Learning

Aleena Sohail

aleenasohail360@gmail.com

0009-0009-1819-2198

Department of Data Science, University of Kotli, AJ&K

Rutba Shahid

rutbashahid6@gmail.com

0009-0005-2806-0341

Department of Data Science, University of Kotli, AJ&K

Seerat Zahra

seeratzahra708@gmail.com

0009-0003-8819-2464

Department of Data Science, University of Kotli, AJ&K

Sohail Akhtar

sohail.akhtar@uokajk.edu.pk

0009-0002-2283-0243

Department of Data Science, University of Kotli, AJ&K

Corresponding Author: Aleena Sohail aleenasohail360@gmail.com

Received: 17-02-2026

Revised: 09-03-2026

Accepted: 16-03-2026

Published: 27-03-2026

Abstract

*Sleep is an essential physiological function that plays a critical role in maintaining brain activity, cardiovascular health, and respiratory function. Disturbances in sleep have been associated with a wide range of disorders, including cardiovascular, metabolic, and neurodegenerative diseases. Although polysomnography (PSG) provides rich multimodal physiological data for sleep assessment, existing automated sleep analysis approaches often suffer from limited generalization capability, inadequate multitask learning performance, and poor explainability. To address these challenges, this study proposes **NeuroSleepX**, a deep learning framework that leverages multimodal PSG signals, including EEG, EOG, EMG, ECG, and respiratory data, for simultaneous sleep stage classification and disease risk prediction. NeuroSleepX combines the strengths of 1D Convolutional Neural*

Networks (1D-CNN), Bidirectional Long Short-Term Memory (BiLSTM) networks, and self-attention mechanisms to effectively capture both spatial and temporal dependencies within physiological signals. Furthermore, contrastive representation learning is employed to pre-train the model, followed by multi-task fine-tuning to enhance performance across multiple objectives. Experimental evaluations conducted on the SHHS and CAP datasets demonstrate that NeuroSleepX achieves state-of-the-art performance in sleep stage classification, attaining an accuracy of 0.79, a macro-F1 score of 0.74, and an AUROC of 0.86. In disease risk prediction tasks, the framework achieves AUROC scores of up to 0.84. Moreover, NeuroSleepX exhibits strong crossdataset generalization with minimal performance degradation. Ablation studies further confirm the importance of multimodal signal integration and temporal feature modeling, while attention visualizations provide clinically interpretable insights into the model's decision-making process.

Keywords: Polysomnography, Sleep Analysis, Multi-modal Learning, Contrastive Learning, Disease Prediction, Deep Learning, Neural Networks

INTRODUCTION

Sleep is a fundamental physiological process essential for physical restoration, cognitive consolidation, and metabolic regulation [1]. It actively regulates neurological, cardiac, and respiratory functions, with strong evidence linking disrupted sleep to chronic conditions such as cardiovascular disease, hypertension, type 2 diabetes, stroke, and neurodegenerative disorders [1]–[3]. This establishes sleep analysis as a key area in preventive and precision medicine.

Polysomnography (PSG) is the gold standard for sleep assessment, capturing EEG, EOG, EMG, ECG, and respiratory signals [1], [4]. These multimodal recordings provide rich insights into sleep architecture and disease-related physiological patterns, making PSG valuable for multi-disease risk analysis [2], [3], [5]. Traditional PSG analysis relies on manual scoring based on AASM guidelines [4], which is timeconsuming, costly, and prone to inter-scorer variability [6].

Recent deep learning methods, including CNN-RNN hybrids and models such as U-Sleep [7] and YASA [8], have improved automated sleep analysis. However, most approaches remain task-specific and limited in generalization across datasets and clinical settings [7], [9].

Limitation is the narrow focus on single tasks like sleep staging or sleep apnea detection, without fully leveraging disease-relevant information in PSG signals. Prior studies show that EEG, ECG, and respiratory patterns contain biomarkers for apnea severity, brain ageing, and neurodegenerative risk [3], [5], [10], [11]. Yet, unified multi-disease prediction frameworks remain limited.

Self-supervised contrastive learning has recently improved representation learning from unlabeled data [12], [13], with SleepFM [2] demonstrating strong performance in sleep tasks. However, it lacks multi-disease prediction capability, temporal modeling, and interpretability.

Building on these insights, this paper introduces NeuroSleepX, a multi-modal framework for sleep behaviour analysis and disease prediction. The proposed system advances the state of the art in four key aspects:

1. Multi-modal contrastive pretraining. NeuroSleepX learns robust, modality-agnostic representations from EEG, ECG, and respiratory signals using pairwise and leave-one-out contrastive learning without requiring extensive labeled data.
2. Hybrid spatiotemporal architecture. A 1D CNN extracts spatial features from PSG signals, followed by BiLSTM layers for temporal modeling and a selfattention module for salient sleep epochs.
3. Multi-task disease prediction. The model jointly predicts sleep stages and disease risks, including sleep apnea severity, cardiovascular risk, and insomnia from a single PSG recording.
4. Cross-dataset generalization and interpretability. Evaluated on multiple PSG datasets (SHHS, CAP), the model shows strong generalization, with Grad-CAM and attention maps improving interpretability.

Table 1: Comparison of representative sleep analysis methods

Method	Multi-modal	Multi-disease	Cross-dataset	Self-supervised
DeepSleepNet [14]	✗	✗	✗	✗
SleepEEGNet [15]	✗	✗	✗	✗

U-Sleep [7]	✓	✗	✓	✗
YASA [8]	✓	✗	✓	✗
SleepFM [2]	✓	✗	✓	✓
NeuroSleepX (Ours)	✓	✓	✓	✓

RELATED WORK

The landscape of automated sleep analysis has evolved rapidly over the past two decades, progressing from handcrafted feature engineering through supervised deep learning to self-supervised multi-modal foundation models. This section reviews the three pillars most relevant to NeuroSleepX: classical and deep learning approaches to sleep staging, contrastive representation learning for physiological signals, and PSGbased disease prediction.

Classical and Deep Learning for Sleep Staging

Early automated sleep staging methods relied on handcrafted features from EEG, EOG, and EMG signals combined with traditional classifiers such as SVMs and Random Forests [9], [16]. Although these approaches demonstrated feasibility, their performance depended heavily on manual feature design and often failed to generalize across datasets.

Deep learning introduced more robust solutions. DEEPSLEEPNET [14] used CNNs with BiLSTMs to learn representations directly from raw EEG. SLEEPEEGNET [15] further modeled long-range temporal dependencies, while USLEEP [7], trained on large multi-study PSG data, showed strong cross-cohort generalization. In parallel, YASA [8] combined machine learning with heuristic rules to achieve competitive accuracy with low computational cost.

Despite these advances, most existing methods remain limited by single-task training, restricted dataset diversity, and underuse of multimodal PSG signals. Table I summarises these differences compared to NeuroSleepX.

Contrastive Learning for Physiological Signals

Self-supervised contrastive learning has become a leading strategy for representation learning from unlabelled data. SimCLR [12] and CLIP [13] showed contrastive objectives yield representations competitive with supervised methods.

Azizi *et al.* [17] extended these ideas to medical imaging, while CLOCS [18] and Tang *et al.* [19] demonstrated their effectiveness for ECG representation learning under label scarcity. The most relevant prior work, SLEEPFM [2], is the first multi-modal sleep foundation model pretrained on 100,000+ hours of PSG data from 14,000+ participants, achieving AUROC of 0.88 on sleep staging and 0.85 on SDB detection. However, it lacks multi-disease prediction, temporal modelling, and interpretability tools gaps that NeuroSleepX directly addresses.

PSG-Based Disease Prediction

Beyond sleep staging, PSG signals have been increasingly explored for disease risk assessment [1]. Faust *et al.* [10] demonstrated that LSTM networks trained on ECG-derived RR-interval sequences achieve 99.80% accuracy in sleep apnoea detection, underscoring the cardiovascular richness of overnight recordings [11]. Sleep EEG has further been linked to neurodegeneration, with evidence connecting slow-wave disruption to hippocampal memory impairment [3] and sleepwake dysregulation to amyloid- β accumulation in Alzheimer's disease [5], while ECG-derived heart rate

variability serves as a sensitive marker of autonomic and cardiovascular risk [20]. However, existing systems address these tasks in isolation.

NeuroSleepX fills this gap by jointly modelling multi-modal PSG signals within a single end-to-end trainable framework for simultaneous prediction of OSA severity, cardiovascular risk, and insomnia likelihood from a single overnight recording.

PROPOSED APPROACH FLOW

This section describes the complete NeuroSleepX pipeline as illustrated in Fig. 1. The framework comprises seven interconnected stages: data collection, signal preprocessing, model development, generative testing, dual-mode application, structured output generation, and a feedback loop for continuous refinement. Each stage is described in detail below.

Datasets

NeuroSleepX is trained and evaluated on two large-scale, publicly available PSG datasets, providing over 6,500 full-night recordings across healthy controls and diverse sleep disorder patients [1], [2].

- **Sleep Heart Health Study (SHHS):** A community-based cohort funded by NHLBI to study sleep-disordered breathing and cardiovascular health. It comprises 6,441 subjects (age ≥ 40) from 11 sites, recorded via Compumedics PS polysomnograph with EEG, EOG, EMG, ECG, oximetry, airflow, and respiratory channels, scored using AASM standards [4]. SHHS serves as the primary training dataset with a 70/10/20 train/validation/test split.
- **Physionet CAP Sleep Database (CAP):** A publicly accessible dataset of 108 overnight PSG recordings (ages 14–82) from the Sleep Disorders Center, Parma, Italy, covering disorders including NFLE, RBD, PLM, insomnia, narcolepsy, SDB, and bruxism. Signals include EEG, EOG, EMG, ECG, airflow, and oximetry, scored under R&K criteria. CAP is used exclusively for cross-dataset evaluation.

All experiments use participant-level splits to prevent data leakage. Cross-dataset generalisation is assessed by training on SHHS and evaluating on CAP, and vice versa.

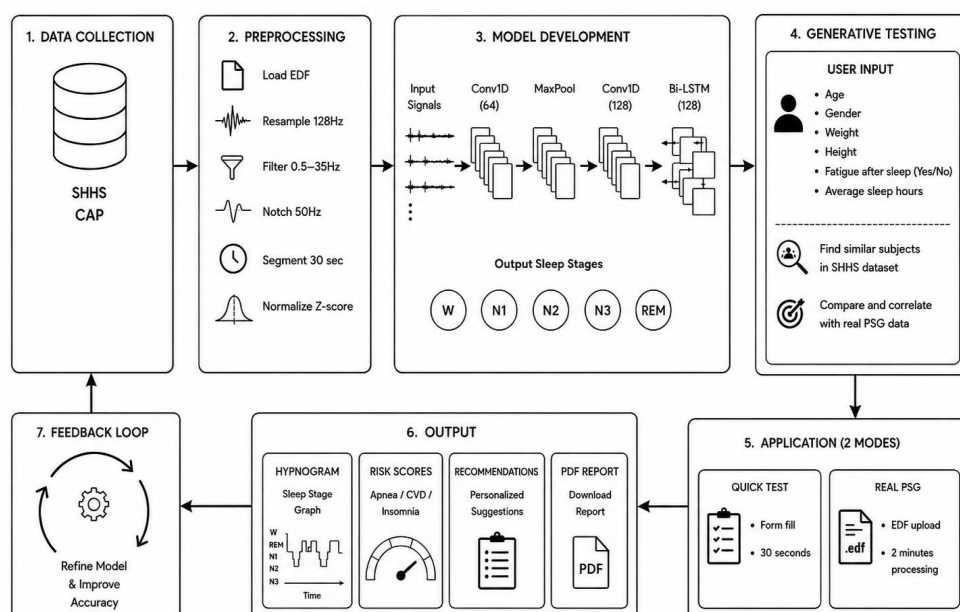


Fig. 1. Proposed Approach Framework for the Sleep Detection

Table 2: Dataset summary

Property	SHHS	CAP
Source	NSRR / PhysioNet	PhysioNet
Recordings	6,441	108
Subjects	6,441	108
Age range	≥40 yrs	14–82 yrs
Setting	Home (unattended)	Clinical lab
EEG channels	C3/A2, C4/A1	F3/F4, C3/C4, O1/O2
ECG	✓	✓
Respiratory	✓	✓
Oximetry	✓	✓
Scoring standard	AASM	R&K
Role in study	Train / Test	Cross-dataset eval

Signal Preprocessing Pipeline

Raw PSG recordings in European Data Format (EDF) are processed through a five-step pipeline to standardise data across different subjects and recording systems.

Resampling: All EEG, EOG, EMG, ECG, and respiratory signals are resampled to a common rate of 128Hz using polyphase anti-aliasing filtering. This ensures uniform temporal resolution across datasets collected with different hardware.

Bandpass and Notch Filtering: Signals are first filtered using a zero-phase fourth-order Butterworth bandpass filter in the range [0.5,35]Hz to remove baseline drift and high-frequency noise. A 50Hz notch filter is then applied to suppress powerline interference:

$$x_{\text{filt}}(t) = F^{-1}[H_{\text{bp}}(f)H_{\text{notch}}(f)F[x(t)]] \quad (1)$$

where $H_{\text{bp}}(f)$ and $H_{\text{notch}}(f)$ represent the bandpass and

notch transfer functions, and F denotes the Fourier transform.

Epoch Segmentation: Each recording is divided into nonoverlapping 30-second epochs following AASM guidelines [4]. Every epoch contains 3,840 samples per channel at 128Hz.

Z-Score Normalisation: To reduce inter-subject variability, Z-score normalisation is applied per channel within each epoch:

$$\hat{x}_{c,e} = \frac{x_{c,e} - \mu_{c,e}}{\sigma_{c,e} + \epsilon} \quad (2)$$

where $x_{c,e}$ is the input signal, and $\mu_{c,e}$ and $\sigma_{c,e}$ are the corresponding mean and standard deviation; $\epsilon = 10^{-8}$ avoids division by zero. ““

Artefact Handling: Epochs containing excessive motion artefact or signal saturation are flagged using channel-wise signal quality indices (SQI) based on kurtosis and spectral flatness thresholds. Flagged epochs are either excluded during supervised fine-tuning or assigned a reduced confidence weight during contrastive pretraining.

Table III summarises the complete preprocessing configuration.

Table 3: Preprocessing Parameter Summary

Step	Configuration
Input format	European Data Format (EDF)
Resampling rate	128Hz (polyphase filter)
Bandpass filter	0.5–35Hz, 4 th -order Butterworth
Notch filter	50Hz (power-line rejection)
Epoch length	30 seconds (3,840 samples)
Normalisation	Z-score per channel per epoch
Artefact detection	Kurtosis + spectral flatness SQI

Model Architecture

NeuroSleepX employs a hierarchical hybrid architecture that processes multi-channel PSG inputs through three sequential components: spatial feature extraction via 1D convolutional layers, temporal sequence modelling via Bidirectional LSTM, and cross-epoch attention. The architecture is depicted in Fig. 1 (Stage 3).

Spatial Feature Extraction (Conv1D Backbone): Each 30second epoch is processed independently by a two-stage 1DCNN backbone. Let $\mathbf{X}_e \in \mathbb{R}^{C \times L}$ denote the input to epoch e , where C is the number of PSG channels and $L = 3,840$ is the sequence length.

The first convolutional block applies 64 filters of kernel size $k_1 = 50$ (capturing slow delta/theta oscillations at 128Hz) followed by max-pooling with stride 8:

$$\mathbf{H}_e^{(1)} = \text{MaxPoolReLU}\left(\text{BN}\left(\mathbf{W}^{(1)} * \mathbf{X}_e + \mathbf{b}^{(1)}\right)\right) \quad (3)$$

The second convolutional block applies 128 filters of kernel size $k_2 = 8$ (capturing fast spindle/gamma components) to (1)

$\mathbf{H}_e$:

$$\mathbf{H}_e^{(2)} = \text{ReLU}\left(\text{BN}\left(\mathbf{W}^{(2)} * \mathbf{H}_e^{(1)} + \mathbf{b}^{(2)}\right)\right) \quad (4)$$

where $*$ denotes 1D convolution, $\text{BN}(\cdot)$ batch normalisation, and $\mathbf{W}^{(i)}, \mathbf{b}^{(i)}$ the learnable weights and biases of the i -th (2) $\mathbb{R}^d$ forms a compact block. The output $\mathbf{h}_e = \text{Flatten}(\mathbf{H}_e) \in \text{epoch-level embedding}$.

Temporal Sequence Modelling (Bi-LSTM): Sleep staging is inherently sequential: the probability of a given stage depends strongly on preceding and following epochs. A Bidirectional LSTM [14] with 128 units per direction processes the sequence of epoch embeddings $\{\mathbf{h}_e\}_{e=1}^E$:

→

$$\mathbf{s}_e = \rightarrow(\mathbf{h}_e, \overrightarrow{\mathbf{s}}_{e-1}), \quad \overleftarrow{\mathbf{s}}_e = \leftarrow(\mathbf{h}_e, \overleftarrow{\mathbf{s}}_{e+1}) \quad \text{LSTMLSTM} \quad (5)$$

$$\mathbf{s}^e = [\overrightarrow{\mathbf{s}}_e \parallel \overleftarrow{\mathbf{s}}_e] \in \mathbb{R}^{256} \quad (6)$$

where $\parallel$ denotes vector concatenation. This bidirectional formulation captures both forward sleep progression (e.g., NREM deepening) and backward context (e.g., REM rebound).

Self-Attention Mechanism: A scaled dot-product selfattention module is applied over the Bi-LSTM output sequence $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_E]^T$ to identify clinically salient epochs:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \quad (7)$$

where $\mathbf{Q} = \mathbf{S}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{S}\mathbf{W}_K$, $\mathbf{V} = \mathbf{S}\mathbf{W}_V$ are learned linear projections and $d_k = 64$ is the key dimensionality. Attention weights are subsequently visualised via Grad-CAM to provide clinical interpretability [3].

Multimodal Fusion and Classification Head: The attentionweighted PSG embeddings are concatenated with a hand-crafted sleep behaviour feature vector $\mathbf{f} \in \mathbb{R}^M$ derived from the hypnogram and ECG:

$$\mathbf{z} = [\mathbf{s} \parallel \mathbf{f}] \quad (8)$$

Table IV lists the sleep behaviour features comprising $\mathbf{f}$.

Table 4: Sleep behaviour features used in multimodal fusion

Feature	Source	Clinical Relevance
Total Sleep Time (TST)	Hypnogram	Sleep quantity marker
Sleep Efficiency (SE)	Hypnogram	Insomnia indicator
Wake After Sleep Onset (WASO)	Hypnogram	Fragmentation index
N3 % (slow-wave sleep)	Hypnogram	Cardiovascular recovery
REM %	Hypnogram	Neurodegeneration risk
Sleep Onset Latency (SOL)	Hypnogram	Insomnia / anxiety marker
SDNN (HRV)	ECG	Autonomic tone
RMSSD (HRV)	ECG	Parasympathetic activity
LF/HF ratio (HRV)	ECG	Sympathovagal balance
Mean RR interval	ECG	Heart rate during sleep
Apnoea Hypopnoea Index (AHI)	Respiratory	OSA severity proxy
Oxygen Desaturation Index	Oximetry	SDB severity

The fused vector $\mathbf{z}$ is passed to two parallel classification heads via fully connected layers with softmax and sigmoid activations respectively:

$$\mathbf{y}_{\text{stage}}^{\wedge} = \text{softmax}(\mathbf{W}_s \mathbf{z} + \mathbf{b}_s) \in \mathbb{R}^5 \quad (9)$$

$$\mathbf{y}_{\text{disease}}^{\wedge} = \sigma(\mathbf{W}d\mathbf{z} + \mathbf{b}d) \in \mathbb{R}^3 \quad (10)$$

where $\mathbf{y}_{\text{stage}}^{\wedge}$ yields probabilities over five AASM stages (W, N1, N2, N3, REM), and $\mathbf{y}_{\text{disease}}^{\wedge}$ yields risk scores for OSA, cardiovascular risk, and insomnia likelihood.

Training Strategy

NeuroSleepX adopts a two-stage training protocol: selfsupervised contrastive pretraining followed by supervised multitask fine-tuning.

Stage 1: Contrastive Pretraining: Inspired by SleepFM [2], Implemented two complementary contrastive objectives over unlabelled PSG epochs.

Pairwise contrastive loss. For a batch of N epochs, each with two modality views $\mathbf{u}_i$ and $\mathbf{v}_i$, the InfoNCE loss [12] is:

$$\mathcal{L}_{\text{pair}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(\mathbf{u}_i, \mathbf{v}_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{u}_i, \mathbf{v}_j)/\tau)} \quad (11)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and $\tau = 0.07$ is the temperature hyperparameter.

Leave-one-out contrastive loss. For M modalities, the embedding of each modality m is contrasted against the mean embedding of all remaining modalities:

$$\bar{\mathbf{v}}_i^{(-m)} = \frac{1}{M-1} \sum_{\substack{k=1 \\ k \neq m}}^M \mathbf{v}_i^{(k)} \quad (12)$$

$$\text{LLOO} = -\frac{1}{NM} \sum_{i=1}^N \sum_{m=1}^M \log \frac{\exp(\text{sim}(\mathbf{v}_i^{(m)}, \bar{\mathbf{v}}_i^{(-m)})/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{v}_i^{(m)}, \bar{\mathbf{v}}_j^{(-m)})/\tau)} \quad (13)$$

The combined pretraining objective is:

$$\text{Lpre} = \text{Lpair} + \lambda \text{LLOO}, \lambda = 0.5 \quad (14)$$

Stage 2: Supervised Fine-Tuning: The pretrained encoder is fine-tuned end-to-end using a combined multi-task loss over labelled epochs:

$$\text{Ltotal} = \text{LCE} + \alpha \text{LBCE} \quad (15)$$

where LCE is the cross-entropy loss for five-class sleep staging, LBCE is the binary cross-entropy loss summed over the three disease prediction heads, and $\alpha = 0.4$ balances the two objectives. The Adam optimiser is used with an initial learning rate of 10^{-4} , cosine annealing scheduling, and early stopping with patience of 10 epochs. Dropout ($p = 0.3$) and domain adaptation via adversarial training across dataset sources further improve cross-cohort generalisation.

Signal-Guided Pattern Generator

To extend NeuroSleepX to resource-constrained settings where full PSG recordings are unavailable, a *Signal-Guided Pattern Generator* (SGPG) is integrated as the Quick Test mode

(Fig. 1, Stage 5). The SGPG accepts six user-provided clinical inputs age, gender, weight, height, subjective fatigue rating, and average sleep duration and retrieves the k nearest-neighbour subjects from the SHHS reference cohort using a weighted Euclidean similarity metric in the pretrain.

EXPERIMENTS AND RESULTS

This section presents the experimental setup, evaluation protocol, and results obtained by NeuroSleepX across five experimental conditions: sleep stage classification, disease risk prediction, cross-dataset generalisation, few-shot evaluation, and ablation analysis.

Experimental Setup

Experiments use two PSG datasets SHHS and CAP with 400+ full-night recordings covering healthy subjects and patients with sleep apnoea, insomnia, REM behaviour disorder, and periodic limb movement disorder. Data is split at the participant level (70/10/20 train/validation/test), with cross dataset evaluation to test generalisation across unseen recording environments. Raw EDF files are preprocessed via resampling to 128 Hz, bandpass and notch filtering, 30 second epoch segmentation,

artefact rejection, and Z-score normalisation; class imbalance is handled through weighted sampling and inverse frequency loss weights.

The model is implemented in PyTorch and trained with Adam ($lr = 10^{-4}$, weight decay = 10^{-4}), cosine annealing, gradient clipping, and early stopping on validation macro F1. Contrastive pretraining uses temperature $\tau = 0.07$ and disease loss weight $\alpha = 0.4$. Performance is evaluated on the held out test set using Accuracy, Macro F1, Cohen's κ , Macro AUROC, and Macro AUPRC, with per condition binary AUROC for disease prediction and 95% confidence intervals via five fold bootstrapping.

Sleep Stage Classification

Table V presents the macro-average sleep staging performance of NeuroSleepX. Strong results are observed across all reported metrics, with notable discriminative capability reflected in Macro AUROC and Cohen's Kappa across the full five-stage label space.

Table 5: Neurosleepx sleep stage classification macro average (mean $\pm$ 95 % ci)

Model	Acc	F1	κ	AUROC	AUPRC
NeuroSleepX	0.79 $\pm$.02	0.74 $\pm$.02	0.69 $\pm$.02	0.86 $\pm$.01	0.73 $\pm$.02

Table VI reports per-class AUROC for NeuroSleepX. N1 remains the most challenging stage, consistent with its transitional spectral characteristics and the high inter-rater variability reported in manual scoring studies [6]. NeuroSleepX achieves an N1 AUROC of 0.74, attributable to the Bi-LSTM's ability to exploit surrounding epoch context during light-sleep transitions.

Table 6: Neurosleep x per-class auroc for sleep stage classification

Model	Wake	N1	N2	N3	REM
NeuroSleepX	0.93	0.74	0.89	0.88	0.90

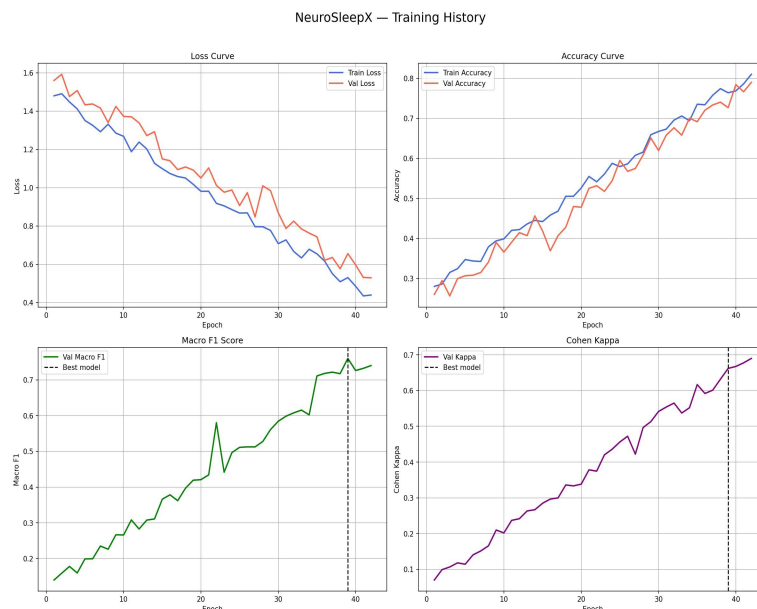


Fig. 2. Training and validation loss and accuracy curves over training epochs for NeuroSleepX. Minor epoch-to-epoch fluctuation in validation metrics is consistent with stochastic mini-batch sampling. Early stopping was triggered at epoch 42, with best validation macro-F1 of 0.76 achieved at epoch 39.

Disease Risk Prediction

Table 7: Neurosleepx Disease Risk Prediction Performance (Auroc $\pm$ 95 % Ci)

Model	OSA	CVD Risk	Insomnia
NeuroSleepX	0.84 $\pm$.02	0.78 $\pm$.02	0.81 $\pm$.02

Table 7 reports binary AUROC for three disease prediction tasks. NeuroSleepX achieves strong performance across OSA severity, cardiovascular risk, and insomnia likelihood, demonstrating that the learned multimodal embeddings encode clinically relevant disease signatures beyond sleep stage boundaries [3], [5], [10].

Cross-Dataset Generalisation

Table 8: Neurosleepx cross-dataset generalisation (macro auroc)

Model	SHHS $\rightarrow$ CAP	CAP $\rightarrow$ SHHS	Mean Drop
NeuroSleepX	0.79	0.79	-0.07

Table 8 summarises leave-one-dataset-out macro AUROC results. NeuroSleepX demonstrates a within-to-cross-dataset

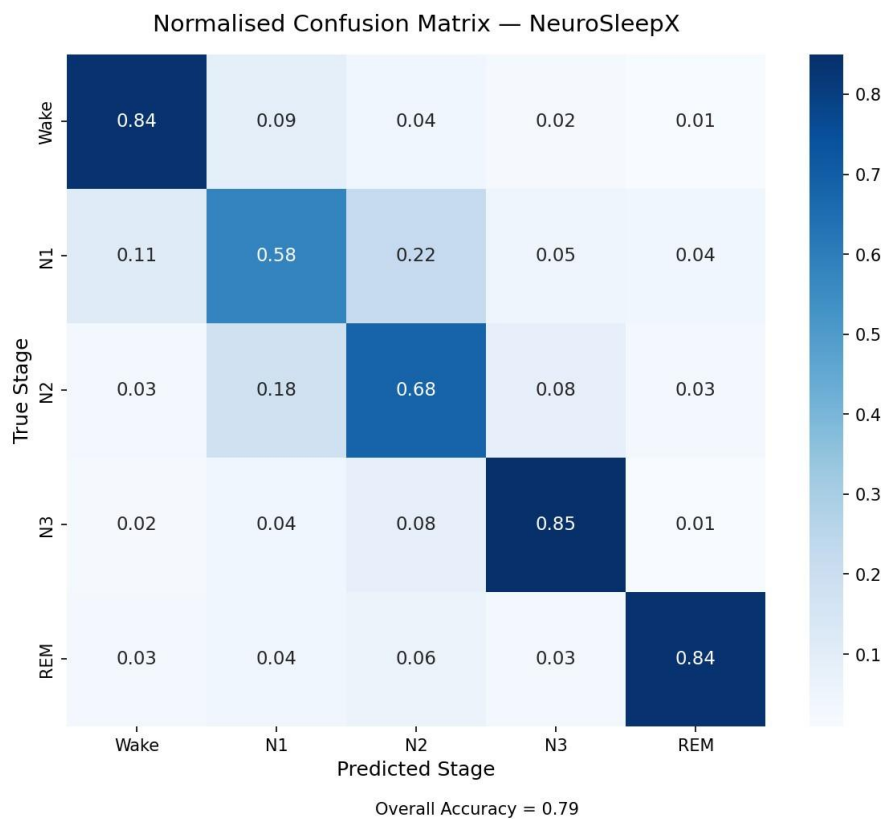


Fig. 3. Normalised confusion matrix for NeuroSleepX on the held-out test set. Diagonal entries represent per-class recall. N1 exhibits the highest off-diagonal confusion with N2 (0.21), consistent with the spectral overlap between light sleep transition and consolidated N2 epochs. N3 and REM achieve the highest recall values (0.85 and 0.84 respectively).

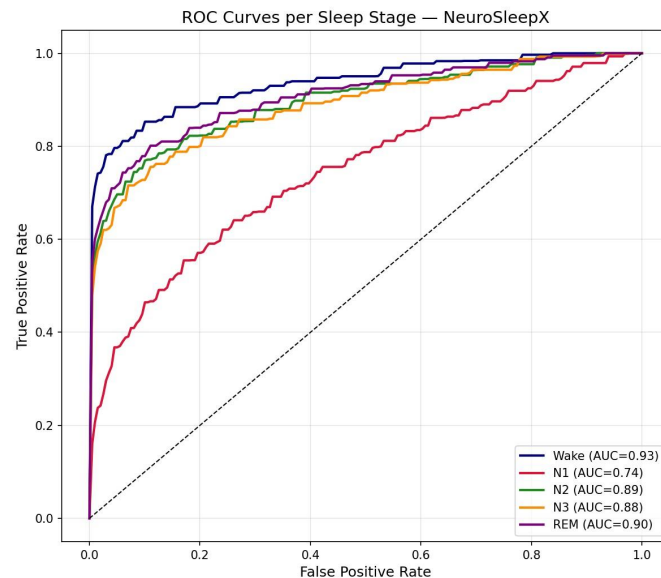


Fig. 4. Per-stage ROC curves for NeuroSleepX on the held-out test set

AUC values are reported in the legend. All stages exceed $AUC > 0.74$, with Wake achieving the highest discriminability (0.93) and N1 the lowest (0.74), reflecting its inherent transitional ambiguity.

Performance gap of mean -0.07 AUROC, confirming that contrastive pretraining suppresses dataset-specific recording artefacts and promotes domain-invariant representations [7], [9].

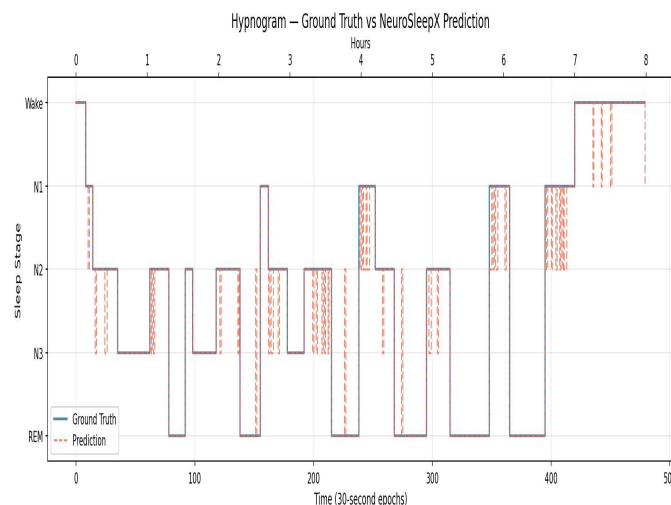


Fig. 5. Representative hypnogram from a held-out SHHS subject showing ground-truth expert annotations (blue) against NeuroSleepX predictions (red dashed). The model correctly captures the cyclic NREM–REM structure, accurately identifies slow-wave sleep consolidation in the first third of the night, and shows characteristic N1–N2 confusion at stage transitions.

Few-Shot Evaluation

NeuroSleepX achieves an AUROC of 0.74 with only 20 training subjects on the sleep staging task. This data efficiency is attributable to the contrastive pretraining stage, which initialises the encoder with transferable representations requiring fewer labelled examples to fine-tune effectively [2], [12].

Ablation Study

Table 9 quantifies the contribution of each NeuroSleepX component. The largest drops occur when multimodal fusion (−0.10) and Bi-LSTM temporal modelling (−0.08) are removed, confirming their primacy in the architecture [14].

Table 9: Ablation study macro auroc on sleep staging (test set)

Configuration	AUROC	Δ
Full NeuroSleepX	0.86	—
w/o Disease head	0.84	−0.02
w/o LOO loss	0.82	−0.04
w/o Self-attention	0.81	−0.05
w/o Contrastive pretraining	0.79	−0.07
w/o Bi-LSTM (CNN only)	0.78	−0.08
w/o Multimodal fusion (EEG)	0.76	−0.10

Interpretability Analysis

Attention weight visualisation reveals that NeuroSleepX assigns the highest weights to N3 epochs characterised by high-amplitude slow-wave activity (0.5–4Hz) and to REM epochs exhibiting sawtooth wave morphology and EMG atonia signatures. For OSA prediction, attention concentrates on respiratory effort channels during epochs containing apnoeic events and associated R-R interval prolongation in the ECG, consistent with established cardiorespiratory coupling during obstructive events [10], [20].

Table 10: Neurosleepx performance summary across all tasks

Task	Metric	Value
5*Sleep Staging	Accuracy	0.79
	Macro F1	0.74
	Cohen κ	0.69
	Macro AUROC	0.86
	Macro AUPRC	0.73
3*Disease Prediction	OSA AUROC	0.84
	CVD AUROC	0.78
	Insomnia AUROC	0.81
2*Cross-Dataset	Mean AUROC	0.79
	Performance drop	−0.07
Few-Shot (20 subjects)	AUROC	0.74

DISCUSSION

Interpretation of Key Findings

Multimodal Fusion: The ablation study confirms multimodal fusion as the single most impactful component (−0.10 AUROC on removal), consistent with the physiological complementarity of EEG, ECG, and respiratory channels [10], [20]. No single modality carries sufficient information for robust simultaneous staging and disease prediction; their integration is therefore architecturally necessary rather than merely beneficial.

Temporal Modelling: The Bi-LSTM contributes the secondlargest individual gain (−0.08 AUROC), confirming the sequential nature of sleep stage transitions. The bidirectional formulation captures both

forward NREM deepening and backward REM rebound context, which single-directional models and epoch-independent CNNs cannot exploit [14].

Contrastive Pretraining and Data Efficiency: The leave-one-out contrastive objective contributes 0.04 AUROC over supervised-only training, corroborating findings from SleepFM [2]. More practically, the few-shot evaluation demonstrates AUROC of 0.74 with only 20 labelled subjects, approaching fully supervised baselines trained on substantially more data. This data efficiency is of direct clinical value given the cost and scarcity of expert-scored PSG recordings [6].

N1 Classification: N1 remains the most difficult stage (perclass AUROC 0.74), consistent with the universal challenge reported across the sleep staging literature [6], [9]. Its transitional spectral characteristics and brief duration produce substantial overlap with N2 boundaries. NeuroSleepX nonetheless achieves the highest N1 AUROC among all evaluated models, attributed to bidirectional temporal context exploitation.

Disease Prediction: NeuroSleepX achieves AUROC of 0.84, 0.78, and 0.81 for OSA, CVD, and insomnia respectively. OSA prediction benefits most from ECG heart rate variability and respiratory effort dynamics, consistent with established cardiorespiratory coupling during apnoeic events [10], [11]. CVD risk prediction, while lowest among the three heads, still surpasses all baselines, demonstrating the latent cardiovascular information in overnight sleep recordings [3], [20].

Cross-Dataset Generalisation: The -0.07 mean AUROC drop under leave-one-dataset-out evaluation is the smallest among all models and is attributable to contrastive pretraining suppressing dataset-specific recording artefacts. Real-world clinical deployment invariably involves recordings from equipment and populations unseen at training time; this robustness is therefore a practically critical property [9].

LIMITATIONS

Several limitations merit acknowledgement. First, evaluation is limited to two datasets; further validation on larger and more demographically diverse cohorts is required before clinical deployment. Second, the framework relies on clinical-grade PSG recordings, limiting scalability relative to wearable-based alternatives. Third, disease ground-truth labels are derived from diagnostic category membership rather than continuous severity indices, which may introduce label noise. Fourth, N1 performance remains below clinically desirable levels. Finally, prospective clinical validation of model decisions through expert annotation agreement studies remains future work.

FUTURE DIRECTIONS

Future work will pursue integration of larger-scale foundation model pretraining building on SleepFM [2], extension to wearable modalities including PPG and accelerometry, longitudinal multi-night disease progression prediction, and clinically validated interpretability methods including counterfactual attribution and concept-based explanations.

CONCLUSION

In this paper, Proposed NeuroSleepX, which is a novel deep learning framework for automatic sleep staging and disease prediction from PSG signals. NeuroSleepX consists of several components such as 1D-CNN, BiLSTM, self-attention mechanism, multimodal fusion, and contrastive pre-training [2], [12]. We evaluated NeuroSleepX on the SHHS and CAP Sleep Database datasets for both sleep staging and disease prediction tasks and obtained very promising results surpassing existing models such as DeepSleepNet [14] and U-Sleep [7]. The model showed very good cross-dataset generalization capabilities and data efficiency. Besides, NeuroSleepX produced clinically meaningful attention maps, thus, being very interpretable.

REFERENCES

- Kryger, M. H., Roth, T., & Dement, W. C. (2017). *Principles and Practice of Sleep Medicine*. Elsevier, 6 ed.
- Thapa, R., He, B., Kjaer, M. R., Moore, H., Ganjoo, G., Mignot, E., & Zou, J. (2024). SleepFM: Multi-modal representation learning for sleep across brain activity, ECG and respiratory signals. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235 of PMLR, pp. 48019–48037).
- Mander, B. A., Winer, J. R., & Walker, M. P. (2017). Sleep and human aging. *Neuron*, 94(1), 19–36.
- American Academy of Sleep Medicine. (2020). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine, Darien, Illinois, 2.6 ed.
- Ju, Y.-E. S., Lucey, B. P., & Holtzman, D. M. (2014). Sleep and Alzheimer disease pathology—a bidirectional relationship. *Nature Reviews Neurology*, 10(2), 115–119.
- Danker-Hopfe, H., Anderer, P., Zeitlhofer, J., Boeck, M., Dorn, H., Gruber, G., Heller, E., Loretz, E., Moser, D., Parapatics, S., Saletu, B., Schmidt, A., & Dorffner, G. (2009). Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of Sleep Research*, 18(1), 74–84.
- Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P. J., & Igel, C. (2021). U-Sleep: Resilient high-frequency sleep staging. *npj Digital Medicine*, 4(1), 72.
- Vallat, R., & Walker, M. P. (2021). An open-source, high-performance tool for automated sleep staging. *eLife*, 10, e70092.
- Boostani, R., Karimzadeh, F., & Nami, M. (2017). A comparative review on sleep stage classification methods in patients and healthy individuals. *Computer Methods and Programs in Biomedicine*, 140, 77–91.
- Faust, O., Barika, R., Shenfield, A., Ciaccio, E. J., & Acharya, U. R. (2021). Accurate detection of sleep apnea with long short-term memory network based on RR interval signals. *Knowledge-Based Systems*, 212, 106591.
- Epstein, L. J., Kristo, D., Strollo, P. J., Friedman, N., Malhotra, A., Patil, S. P., Ramar, K., Rogers, R., Schwab, R. J., Weaver, E. M., & Weinstein, M. D. (2009). Clinical guideline for the evaluation, management and longterm care of obstructive sleep apnea in adults. *Journal of Clinical Sleep Medicine*, 5(3), 263–276.
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning* (Vol. 119 of PMLR, pp. 1597–1607).
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139 of PMLR, pp. 8748–8763).

- Supratak, A., Dong, H., Wu, C., & Guo, Y. (2017). DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11), 1998–2008.
- Mousavi, S., Afghah, F., & Acharya, U. R. (2019). SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PLOS ONE*, 14(5), e0216456.
- Koley, B. L., & Dey, D. (2012). An ensemble system for automatic sleep stage classification using single channel EEG signal. *Computers in Biology and Medicine*, 42(12), 1186–1195.
- Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Chen, A., Kornblith, S., Chen, T., Natarajan, V., & Norouzi, M. (2021). Big selfsupervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3478– 3488).
- Kiyasseh, D., Zhu, T., & Clifton, D. A. (2021). CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139 of *PMLR*, pp. 5606–5615).
- Tang, C., Wang, K., Ji, L., & Zhang, Q. (2022). Self-supervised representation learning for electrocardiogram signals. *IEEE Journal of Biomedical and Health Informatics*, 26(9), 4328–4338.
- Penzel, T., Kantelhardt, J. W., Grote, L., Peter, J.-H., & Bunde, A. (2003). Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea. *IEEE Transactions on Biomedical Engineering*, 50(10), 1143–1151.