applied sciences

MDPI

Article

Hierarchical Multimodal Sleep Staging with Optimized EEG, EOG, and PPG Features for Wearable Applications

Roberto De Fazio [1,2,*], Matteo Paiano [1], Ramiro Velazquez [2], Carolina Del-Valle-Soto [3] and Paolo Visconti [1,2]

[1] Department of Innovation Engineering, University of Salento, Road to Monteroni, 73100 Lecce, Italy; matteo.paiano@unisalento.it (M.P.); paolo.visconti@unisalento.it (P.V.)
[2] Facultad de Ingeniería, Universidad Panamericana, Aguascalientes 20296, Mexico; rvelazquez@up.edu.mx
[3] Facultad de Ingeniería, Universidad Panamericana, Zapopan 45010, Mexico; cvalle@up.edu.mx
* Correspondence: roberto.defazio@unisalento.it; Tel.: +39-0832-297334

Abstract

Automatic sleep staging is fundamental for diagnosing sleep disorders and enabling long-term sleep monitoring with wearable devices. Although Deep Learning has significantly improved classification performance, balancing accuracy with computational efficiency remains challenging, particularly for resource-constrained systems. This paper proposes a lightweight two-stage Deep Learning framework for five-class sleep staging based on optimized multimodal physiological features extracted from electroencephalogram (EEG), electrooculogram (EOG), and photoplethysmography (PPG) signals. The framework is trained and tested using the Bitbrain Open Access Sleep (BOAS) database, considering a 31-subject dataset partitioned into training (24 subjects) and independent test (7 subjects) sets. Feature selection is performed using the minimum Redundancy Maximum Relevance (mRMR) algorithm, followed by Principal Component Analysis (PCA) for EEG and EOG features, while respiratory and cardiac features derived from PPG are directly incorporated into the multimodal representation. A hierarchical Long Short-Term Memory (LSTM) architecture first classifies sleep into Wake, REM, and NREM, then further distinguishes the N1, N2, and N3 stages. On an independent test set, the classifier achieves 88.2% five-class accuracy on the multimodal feature set (EEG + EOG + PPG) with a model size of 3.14 MB, and 87.1% accuracy on the EEG-only feature set using only 2.95 MB of memory. Leave-One-Subject-Out (LOSO) cross-validation yields 86.9% accuracy, supporting subject-independent generalization. Inference latency ranged from 2.55 ms (EEG-only) to 4.02 ms (multimodal), with measured energy per inference of 2.88–10.9 mJ across feature sets. Additional validation on the RichSleep and ISRUC datasets demonstrates robustness across different recording conditions, achieving mean accuracies of 78.2% and 77.3%, respectively. The proposed framework provides a favorable trade-off among classification performance, complexity, and memory footprint, suggesting its potential suitability for wearable and edge-based sleep-monitoring systems.

Keywords: sleep scoring; EEG; EOG; PPG; feature selection; two-stage DL algorithm; LSTM-hierarchical classifier

Academic Editor: Alexander Zhbanov

Received: 4 July 2026
Revised: 5 August 2026
Accepted: 11 August 2026
Published: 16 August 2026

1. Introduction

Sleep is a fundamental physiological process that plays a critical role in brain function, emotional regulation, and overall physical and mental health. Sleep deprivation has been shown to impair cognitive performance, disrupt mood, and negatively impact general

well-being [1–3]. Consequently, effective sleep monitoring is essential for the prevention and early detection of sleep-related disorders. Polysomnography (PSG) is the standard diagnostic method for assessing sleep and associated disorders, involving the simultaneous monitoring of multiple physiological signals [4,5].

Among the various bio-signals recorded during PSG, the electroencephalogram (EEG) remains the most informative modality for classifying sleep stages, diagnosing sleep disorders, and evaluating sleep quality [6,7]. The sensorized sleep mask, developed within a research project funded by the Italian Space Agency (ASI), acquires EEG and other biosignals via non-invasive sensors and enables automated sleep staging using Machine Learning/Deep Learning (ML/DL) algorithms [8]. EEG signals are usually categorized into frequency bands associated with specific physiological and cognitive states. Delta waves, ranging from 0.5 to 4 Hz, are the lowest-frequency brain waves linked to deep sleep, particularly stage N3. Theta waves, with a frequency of 4–8 Hz, are associated with light sleep (stage N1) and states of meditation and relaxation. Alpha waves, ranging from 8 to 13 Hz, are predominant during relaxed wakefulness, especially when the eyes are closed. Beta waves, with frequencies between 13 and 30 Hz, are associated with an active, wakeful state, concentration, and logical thinking. In contrast, gamma waves, with frequencies above 30 Hz, are the highest-frequency brain waves associated with complex cognitive processes, perception, and consciousness [9].

In addition, the electrooculogram (EOG) signals are useful for classifying sleep stages because they reflect changes in eye movements associated with alertness. As drowsiness increases, the frequency of blinks tends to decrease while their duration increases, indicating the transition from wakefulness to sleep [10]. In particular, during the falling asleep phase and the N1 stage, slow eye movements (SEM) appear, characterized by low-frequency oscillations. These phenomena are reflected in increased EOG signal power in the low-frequency bands (delta and theta) and reduced power in the high-frequency bands (beta and gamma), making EOG spectral features effective indicators of sleepiness and the early stages of sleep [11]. Also, the respiratory signal provides relevant information for automatic sleep staging because the respiratory rhythm varies across stages. In particular, breathing rate (BR) tends to decrease and become more regular during Non-Rapid Eye Movement (NREM) sleep, while it is more irregular during Rapid Eye Movement (REM) sleep and wakefulness [12]. The Inspiratory-to-Expiratory Ratio (IER) is a temporal parameter of the respiratory cycle [13] that reflects the dynamics of inspiration and expiration. Studies on respiratory variability show that breathing is more regular and stable in NREM sleep and more variable during REM sleep and wakefulness [14].

During sleep, heart rate (HR) and heart rate variability (HRV) change in characteristic ways between different stages. HR tends to decrease, and HRV increases during deep stages of NREM sleep, while during REM sleep and wakefulness, HR may increase and HRV becomes more variable. This information is extracted from electrocardiogram (ECG) signals recorded via chest or wrist electrodes. Common automatic classification features include HR and HRV indices such as standard deviation of NN interval (SDNN), Root Mean Square of Successive Differences (RMSSD), and low-/high-frequency (LF/HF) ratio, which provide objective indicators for distinguishing wakefulness, NREM, and REM sleep, making them essential for automatic sleep scoring [15].

Accurate sleep staging depends heavily on extracting relevant features from the previously described signals, as these directly influence the performance of classification algorithms. Using an excessive number of features, or features with low discriminative power, increases computational load and noise, potentially degrading model performance. Therefore, targeted feature selection is essential for improving sleep-stage classification and reducing discrepancies between automated and expert manual staging [16]. This study se-

lected and extracted features from physiological signals in the BOAS (Bitbrain Open Access Sleep) dataset, including EEG, EOG, and photoplethysmography (PPG). These signals were recorded with the Brain Quick Plus Evolution PSG system (Micromed). For EEG, a bipolar frontal derivation (F4-F3) was considered, whereas for EOG, a single unipolar derivation was used. Lastly, PPG was processed to derive both cardiac and respiratory features. Based on the scientific literature, we selected the most relevant features for their high information content and discriminative power across the different stages of sleep. In particular, feature sets comprising time-, frequency-, and non-linear-domain features from biosignals were chosen, enabling reliable and meaningful classification models by combining multiple domains. This study then extracted features after preprocessing each signal and applied feature selection and optimization techniques, such as the Minimum Redundancy Maximum Relevance (mRMR) algorithm and Principal Component Analysis (PCA), to EEG and EOG features, selecting the most relevant and least redundant features and yielding compact, information-rich feature sets by combining features from different biosignals.

The reduced feature set was subsequently employed to train and evaluate a classification model for sleep staging. Long Short-Term Memory (LSTM) networks were selected for their proven capability to capture sequential patterns in physiological time-series data. A two-stage hierarchical classification strategy was also implemented, with each model comprising LSTM and dense layers; a first model initially discriminates Wake (W), REM (R), and NREM (N) epochs, followed by a second classifier that refines NREM into N1, N2, and N3 stages. Both five-class and three-class models were evaluated on three feature sets, demonstrating robust classification performance despite the reduced dimensionality.

The main contributions and novelties of the proposed work are:

- A unified pipeline combining mRMR feature selection, PCA dimensionality reduction, data standardization, and a lightweight two-stage LSTM-based framework for resource-constrained deployment. On the 31-subject dataset from the BOAS database (24 for training, seven for testing), the pipeline achieves 87.1% and 88.2% total accuracy on five-class classification for EEG-only and multimodal (EEG + EOG + PPG), respectively, with a memory footprint of 2.95 MB and 3.14 MB. These results suggest that a minimal frontal sensor configuration can deliver competitive sleep-staging performance with low resource demands.
- Comprehensive characterization of the proposed five-class classifier in terms of inference time, latency, and energy per inference across the three feature sets.
- Ablation studies on class-balancing strategies (no balancing, augmentation only, weighted loss only, and augmentation + weighted loss) and loss function formulations (standard cross-entropy, focal loss variants γ = 1.0, 2.0, and transition-aware loss).
- Leave-One-Subject-Out (LOSO) cross-validation, yielding 86.9% $\pm$ 2.9% accuracy and confirming subject-independent generalization capability.
- Cross-dataset validation on two public datasets (RichSleep and ISRUC), achieving mean accuracies of 78.2 and 77.3%, confirming the framework's robustness to domain shift in recording hardware, electrode placement, and population characteristics.
- Comparative analysis against state-of-the-art lightweight architectures, showing that the proposed framework outperforms cascaded TinySleepNet (87.1% accuracy) and SeqSleepNet (84.5% accuracy) under the same evaluation protocol.

The article remainder is organized as follows: Section 2 explores the scientific literature on signal processing and sleep-staging algorithms. Section 3 presents the selection, extraction, and optimization of multi-signal features to obtain reduced-dimensionality datasets for training and testing two-stage DL models for five-class sleep classification (Section 4). Finally, Section 5 discusses the performance of sleep staging models trained

and tested with multi-signal feature sets, compares them with the literature, and reports subject-independent validation on subjects not involved in the prior training/testing phase.

2. Related Works

In recent years, the field of automatic sleep classification has seen substantial advancements, particularly in the application of DL and signal processing techniques to EEG, EOG, ECG, and respiratory data. Increasing attention has been given to frontal and forehead EEG derivations due to their suitability for wearable and ambulatory applications. This section reviews representative works on sleep staging and feature extraction, spanning both DL and traditional Machine Learning (ML) approaches.

A wide range of studies has explored automatic sleep stage classification using a single or a limited number of EEG channels, with growing interest in frontal and forehead derivations due to their practical suitability for wearable applications. In ref. [17], the authors proposed a Deep Learning architecture trained on the Sleep-EDF dataset using the Fpz-Cz derivation, achieving 83.78% accuracy in five-class classification. Their approach demonstrated that high-level features can be effectively extracted even from a single frontal channel. Similarly, ref. [18] focused entirely on forehead EEG channels (Fp1-Fp2) in an ambulatory setting and demonstrated that sleep staging accuracy remained high (79.7–89.1%) even in sleep-deprived subjects, emphasizing the reliability of these accessible channels for non-laboratory applications. Extending DL-based approaches, ref. [19] proposed a Temporal Convolutional Network (TCN) trained on raw single-channel EEG signals (Fpz-Cz), further enhanced by a novel data augmentation technique, and achieved an accuracy of 85.39%. Their results suggest that temporal modeling improves stage transition detection and generalization. Similarly, ref. [20] evaluated a sleep staging algorithm based on a single Fpz-Cz derivation from the Expanded Sleep-EDF dataset. By incorporating feature selection and using Random Forests, they achieved 83.56% accuracy with 57 features, highlighting the importance of informative feature selection.

In another comprehensive study [21], a combination of Random Forests and Hidden Markov Models was applied to datasets including Sleep-EDF, DREAMS, and ISRUC. They used multiple derivations (e.g., Fpz-Cz, Pz-Oz, Cz-A1, and Fp1-A2) to train and test the proposed sleep staging model for five- and six-sleep-stage classifications, achieving accuracies ranging from 77.6% to 87.4%, depending on the dataset and classification scheme, American Academy of Sleep Medicine (AASM) or Rechtschaffen and Kales (R&K). This study highlights both model adaptability and variability introduced by datasets and electrode configurations. More advanced methods have utilized ensemble or decomposition-based architectures; in ref. [22], the authors introduced a hybrid model combining Ensemble Empirical Mode Decomposition (EEMD) with XGBoost and applied it to the Sleep-EDF, DREAMS, and SHHS (Sleep Heart Health Study) datasets. Using derivations such as Pz-Oz, Cz-A1, and C4-A1, they achieved accuracies exceeding 83% across all datasets, with the best result of 91.9% for the Sleep-EDF (five-class) dataset. These findings support the effectiveness of combining signal decomposition with ensemble learners. In ref. [23], the authors reported improved performance using the Null Space Pursuit (NSP) decomposition algorithm on a single-channel EEG (Pz-Oz), achieving accuracies of 93.59% (four-class) and 92.98% (five-class) on the Sleep-EDF dataset, and similarly high values on the DREAMS and SHHS datasets. Their work showcases the power of signal transformation techniques in extracting latent features for accurate classification. In ref. [24], the authors focused on low-complexity REM detection using a frequency-based spectral-edge method on Fp1-A2 and Cz-A1 derivations, achieving an accuracy of 88.52%. This approach was shown to be suitable for low-power and edge devices.

In refs. [25,26], the authors implemented Relevance Vector Machines (RVMs) and hierarchical classifiers using two forehead EEG channels (Fp1 and Fp2) on a custom dataset,

achieving 77% accuracy in five-class classification. These works emphasize that even simplified frontal montages can retain sufficient information when properly modeled. In ref. [27] validated sleep staging using only two forehead electrodes (Fp1-Fp2) in healthy adults and achieved between 80.8% and 85% accuracy, distinguishing five stages. This study demonstrates the feasibility of minimal-channel wearable EEG systems. In a recent study [28], the author compared four Deep Learning architectures (ResNet50-1D, VGG19-1D, Xception-1D, and LSTM) for automatic sleep classification using EEG time-series data from the BOAS dataset. Strong performance was obtained across the evaluated architectures, with the ResNet50-1D model achieving the highest validation accuracy (95%), followed by Xception-1D (94%), VGG19-1D (91%), and LSTM (88%). Yilmaz et al. [29] worked with the same dataset, using EEG features for sleep staging with a Support Vector Machine (SVM) model. They trained a separate binary SVM for each of the five sleep stages and used the continuous distance outputs from SVMs as inputs to a Random Forest (RF) classifier, achieving 87% classification accuracy on the BOAS dataset.

Beyond EEG-only approaches, recent research has shifted toward multimodal frameworks that integrate additional physiological signals, such as EOG, ECG, and respiratory measurements, to improve classifier accuracy. Ref. [30] applied a Convolutional Neural Network (CNN)-LSTM model to detect sleep stages using an EEG derivation (F4-M1), achieving 82.9% accuracy, which increased to 83.8% when combined with an EOG derivation (E1-M2), from a clinical dataset. In ref. [31], the authors developed a 1D-CNN for sleep stage classification using EEG and EOG signals and tested it on Sleep-EDF and Sleep-EDFx datasets. For five-class staging, combining EEG and EOG (91.22% accuracy) outperformed EEG alone (90.8% accuracy), demonstrating the advantage of multimodal inputs. In a recent study, Rashidi et al. used ensemble learning for automatic sleep staging, combining features extracted from EEG sub-bands, HRV, and ECG-Derived Respiration (EDR) [32]. Using a bagging ensemble of decision tree classifiers with 10-fold cross-validation, the authors achieved a mean five-class sleep staging accuracy of 95.09% on the MIT-BIH dataset. When evaluated on the Sleep-EDF dataset, an accuracy of 86.31% was obtained using only a single EEG lead (Fpz-Cz), highlighting that classification performance improves when integrating additional physiological signals. Ref. [33] presented a CNN-LSTM-based sleep staging method that relies solely on cardiovascular and respiratory signals, demonstrating that combining ECG and abdominal respiration achieves strong sleep staging performance on the SHHS dataset. The model reached a Cohen's kappa of approximately 0.58 for EEG-based expert sleep staging, corresponding to 92.9% of the inter-rater agreement typically reported among human EEG scorers. In ref. [34], the authors used an RF and a Multilayer Perceptron (MLP) classifier on the MGH (Massachusetts General Hospital) dataset. Features were extracted from EEG, ECG, electromyography (EMG), and respiratory signals, with the best accuracy (93.72%) achieved by RF using EEG and respiratory signals, showing the benefit of multimodal inputs.

Table 1 summarizes these studies, considering datasets, classification methods, biosignals, EEG leads, targeted sleep stages, and achieved accuracy. In all studies, the EEG electrode position plays a crucial role in determining model performance and usability. While classical derivations such as Pz-Oz or Cz-A1 are common, a shift toward frontal and forehead electrode positions (Fpz-Cz, Fp1, Fp2) is evident, offering better applicability in ambulatory and wearable settings. Moreover, the transition from traditional ML to DL, particularly when combined with advanced preprocessing techniques (e.g., NSP, EEMD), has significantly improved classification accuracy. The best-performing models now exceed 90% accuracy in four- and five-class sleep staging [23,29,32,34]; however, they have come at the cost of high memory and computational requirements, and variability remains across datasets, scoring schemes, and sleep stage distributions.

Table 1. Comparison of the scientific works analyzed in Section 2 regarding algorithms for sleep staging.

Reference	Dataset	ML/DL Algorithms	Number of EEG Leads	Other Biosignals	Classified Sleep Stages	Accuracy
M. Fu et al. [17]	Sleep-EDF dataset, DREAMS subjects dataset	AT-BiLSTM (Attention-based BiLSTM)	1 (Sleep-EDF dataset: Fpz-Cz or Pz-Oz; DRM-SUB dataset: Cz-A1 or Fp1-A1 or O1-A1)	-	5 sleep stages (W, REM, N1, N2, N3)	83.78% (Sleep-EDF Fpz-Cz), 81.72% (DREAMS Fp1-A1)
A. Leino et al. [18]	AES (Ambulatory EEG System) dataset (n = 135) [(a)]	Neural network	1 (Fp1-Fp2)	-	5 sleep stages (W, REM, N1, N2, N3); 4 stages (W, REM, light sleep, deep sleep); 3 stages (W, NREM, REM)	79.7% (5-stage), 84.1% (4-stage), 89.1% (3-stage)
E. Khalili et al. [19]	Sleep-EDF-2013, Sleep-EDF-2018	CNN + temporal; CNN + Conditional; Random Field (CRF) with novel data augmentation	1 (Fpz-Cz or Pz-Oz)	-	5 sleep stages (W, REM, N1, N2, N3)	85.39% (EDF-2013), 82.46% (EDF-2018)
S. Zhao et al. [20]	Expanded Sleep-EDF	SVM, DT, RF, and BPNN with feature selection	1 (Fpz-Cz)	-	5 sleep stages (W, REM, N1, N2, N3)	75.82–83.56% (with 57 features), 75.88–82.53% (with 11 features); in both cases, RF is the best model
H. Ghimatgar et al. [21]	Sleep (S)-EDF, ES-EDF, DREAMS subjects (DRM-SUM), ISRUC sleep datasets	Random Forest + Hidden Markov model	1 (S-EDF, ES-EDF datasets: Fpz-Cz or Pz-Oz; DRM-SUB dataset Cz-A1 or FP1-A2 or O1-A2; ISRUC dataset: C3-A2 or C4-A1	-	6 stages (W, REM, S1, S2, S3, S4) (S-EDF and ES-EDF); 5 stages (W, REM, N1, N2, N3) (DRM-SUM & ISRUC)	79.4–87.4% for six-stage (R&K), 77.6–80.4% for five-stage (AASM) classification

Table 1. *Cont.*

Reference	Dataset	ML/DL Algorithms	Number of EEG Leads	Other Biosignals	Classified Sleep Stages	Accuracy
C. Liu et al. [22]	Sleep-EDF, DREAMS Subjects, SHHS Datasets	EEMD + XGBoost	1 (Sleep-EDF dataset: Pz-Oz; DREAMS dataset: Cz-A1, SHHS dataset: C4-A1	-	5 classes (W, S1, S2, DS, REM); 4 classes (W, LS-combined S1 and S2, DS, REM)	93.1%, 91.9% (Sleep-EDF/4-class, 5-class); 86.4%, 83.4% (DREAMS 4-class, 5-class); 87.5%, 85.8% (SHHS 4-class, 5-class)
W. Xiao et al. [23]	Sleep-EDF, DREAMS Subjects, SHHS Datasets	Null Space Pursuit (NSP) decomposition algorithm	1 (Sleep-EDF dataset: Pz-Oz; DREAMS dataset: Cz-A1; SHHS dataset C4-A1)	-	5 classes (W, N1, N2, N3, REM); 4 classes (W, LS-combined N1 and N2, N3, REM)	93.59%, 92.89% (Sleep-EDF/4-class, 5-class), 91.32%, 90.0% (DREAMS 4-class, 5-class), 90.25%, 88.37% (SHHS 4-class, 5-class)
SA Imtiaz et al. [24]	Custom dataset, PhysioNet Sleep-EDF	Spectral Edge frequency-based method	1 (Fp1-A2 (highest) or Cz-A1 or O1-A2)	-	REM detection	88.52%
C-S. Huang et al. [25]	Custom dataset	Relevance Vector Machine (RVM)	2 (FP1 and FP2)	-	5 sleep stages (W, REM, N1, N2, N3)	76.7 $\pm$ 4.0 (SD) %
C-S. Huang et al. [26]	Custom dataset	Feature selection methods and SVM	2 (FP1 and FP2)	-	Wake, REM, light sleep (N1 and N2), and deep sleep (N3)	77%
D. Popovic et al. [27]	Custom dataset	FP-STAGER algorithm	1 (Fp1-Fp2)	-	5 sleep stages (W, REM, N1, N2, N3)	80.8–85% (for sleep-deprived and well-rested volunteers)
N.P. Duc [28]	BOAS dataset	ResNet50-1D, VGG19-1D, Xception-1D, and LSTM	6 (F3, F4, C3, C4, O1 and O2)	-	5 sleep stages (W, REM, N1, N2, N3)	95% (ResNet50-1D), 91% (VGG19-1D), 94% (Xception-1D), 88% (LSTM) on validation accuracy

Table 1. *Cont.*

Reference	Dataset	ML/DL Algorithms	Number of EEG Leads	Other Biosignals	Classified Sleep Stages	Accuracy
G. Yilmaz et al. [29]	BOAS dataset	SVM + RF	1 (Cz-Oz or Cz-Fz or F3-F4)	-	5 sleep stages (W, REM, N1, N2, N3)	Best accuracy: 87%, for Cz-Fz
H. Korkalainen et al. [30]	Clinical dataset (n = 91) [a]	CNN + LSTM hybrid	1 (F4-M1)	EOG (E1-M2)	5 sleep stages (W, REM, N1, N2, N3)	82.9% (only EEG), 83.8% (EEG + EOG)
O. Yildirim et al. [31]	Sleep-EDF and Sleep-EDFx datasets	1D-CNN	2 (Fpz-Cz and Pz-Oz)	EOG	5 sleep stages (W, REM, N1, N2, N3)	90.8% (only EEG), 91.22% (EEG + EOG)
S. Rashidi et al. [32]	MIT-BIH and Sleep-EDFx datasets	Bagging with Decision tree classifier	1 (Fpz-Cz)	ECG (for HRV and EDR extraction)	5 sleep stages (W, REM, N1, N2, N3)	95.09% (MIT-BIH), 86.31% (Sleep-EDFx)
H. Sun et al. [33]	SHHS dataset	CNN-LSTM	-	ECG and abdominal respiratory signals	5 sleep stages (W, REM, N1, N2, N3)	0.58 Cohen's kappa with respect to EEG-based expert sleep staging, corresponding to 92.9% accuracy
A.M. Tăutan [34]	MGH dataset	RF, MLP	1 (F3-M2)	ECG, EMG, and respiratory signals	5 sleep stages (W, REM, N1, N2, N3)	93.72% (achieved by RF using EEG and respiratory signals)

[a] n: the number of subjects included in the dataset.

3. Materials and Methods

3.1. Description of the BOAS Dataset

In the proposed research activity, we have chosen the Bitbrain Open Access Sleep (BOAS) dataset to extract features from sleep-recorded signals, which were used to train and test the sleep staging algorithm. This dataset was gathered within a project aimed at bridging the gap between gold-standard clinical sleep monitoring using the Micromed Brain Quick Plus Evolution PSG system and emerging wearable EEG technologies, specifically the Bitbrain EEG headband (https://openneuro.org/datasets/ds005555/versions/1.0.0, accessed on 10 November 2025). The dataset includes 128 nights of simultaneous recordings from both systems collected from 108 healthy participants. The PSG system provides a comprehensive, clinically validated set of sleep parameters, while the Bitbrain wearable EEG headband offers a user-friendly, self-administered alternative limited to forehead EEG electrodes. Simultaneous data acquisition enables direct comparison and validation of the wearable EEG device against the established PSG standard, creating a valuable resource for evaluating the performance and potential of wearable EEG technology in sleep studies. For the present study, a subset of 31 subjects was selected from the BOAS database. The selection was restricted to recordings that provided the complete set of physiological signals required by the proposed framework (EEG, EOG, and PPG) and satisfied the adopted signal-quality criteria for reliable feature extraction. Among the recordings that fulfilled these criteria, a subset was randomly selected. The demographic characteristics of the selected subjects are summarized in Table 2.

Table 2. Demographic characteristics of the complete BOAS cohort and of the participants included in the study subset.

Variable	Complete BOAS Cohort (n = 108)	Selected Subset (n = 31)
Age (years)	42.0 ± 18.8	38.0 ± 15.5
Male, n (%)	50 (46.3%)	10 (32.3%)
Female, n (%)	58 (53.7%)	21 (67.7%)
BMI (kg/m^2)	23.8 ± 3.3	24.8 ± 3.2

A rigorous labeling process was adopted to ensure reliable sleep staging. Three expert sleep scorers independently annotated the PSG recordings according to the AASM criteria, using 30 s epochs [35]. A fourth expert derived consensus labels from these annotations to address the inherent variability in human sleep staging, which has an estimated inter-scorer agreement of approximately 85% [36,37]. In detail, the stages are coded as follows: “0” for the Wake stage, “1” for NREM sleep stage 1 (N1), “2” for theNREM sleep stage 2 (N2), “3” for the NREM sleep stage 3, and “4” for REM stage, while if a 30 s epoch contains a PSG disconnect, it is labeled “8”; if the epoch is affected by artifacts, it is labeled “−2”. These consensus labels were then applied to the corresponding wearable EEG recordings, leveraging the simultaneous data acquisition. A CNN model was also used to analyze the dataset [38]. Using a cross-validation procedure, the model was trained and validated separately on the PSG and wearable EEG datasets. It achieved 87.08% agreement between human-consensus labels and model-predicted labels for the PSG data and 86.64% for the wearable EEG data. These results highlight the potential of wearable EEG technology to approximate the accuracy of traditional PSG systems, offering a promising alternative for sleep monitoring in clinical and non-clinical settings. The study highlights the importance of rigorous validation and consensus labeling in ensuring the reliability of sleep staging and the effectiveness of ML models in advancing sleep research.

3.2. Feature Extraction and Analysis for Sleep Staging and Detecting Related Disorders

Accurate sleep staging relies on the analysis of multiple physiological signals recorded during sleep, including EEG, EOG, and PPG-derived information, each capturing complementary aspects of sleep physiology. Specifically, EEG reflects changes in brain activity across sleep stages; EOG represents ocular movements; and PPG provides indirect information on respiratory and cardiac dynamics. From these signals, features are extracted in the time, frequency, and nonlinear domains to capture physiologically meaningful patterns associated with different sleep states. Feature selection plays a critical role in achieving high sleep staging performance, as redundant or weakly discriminative features increase computational complexity and may degrade classification performance. For this reason, a literature analysis has been conducted, examining scientific works on multi-signal analysis for sleep staging and the identification of related disorders. Based on this analysis, Table 3 reports EEG features extracted in the time-, frequency-, and nonlinear- domains, together with their physiological significance. Table 4 summarizes EOG features in the time and frequency domains, highlighting correlation with the sleep stage. Table 5 presents time-domain PPG features related to respiratory activity, while Table 6 reports PPG features in both the time and frequency domains associated with cardiac activity.

The training and test sets were obtained by processing EEG, EOG, and PPG signals from 31 unique subjects in the BOAS dataset, totaling 230 h of sleep and corresponding to 165636 5 s sub-epochs. As described in Section 3.1 below, the 31 subjects were partitioned into training (24 subjects) and testing (7 subjects) sets used to develop and validate the proposed two-stage LSTM classifier. Since the BOAS dataset provides sleep stage labels at 30 s resolution (following the AASM manual), each 5 s sub-epoch inherited the label of its parent 30 s epoch. It is important to clarify that the ground-truth annotation resolution remains 30 s; the 5 s segmentation does not increase the temporal resolution of the labels. Instead, it provides finer temporal sampling of the physiological features and enables the construction of longer input sequences (10×5 s = 50 s) for the LSTM model. This approach is consistent with prior DL-based sleep staging studies that employ sub-epoch segmentation for feature extraction and temporal sequence modeling. Further details on the choice to work on 5 s sub-epochs are provided in Section 3.5. Of the BOAS dataset's 108 participants, 31 were selected for this study based on the simultaneous availability of all required signals: frontal EEG (F4-F3 lead), single-channel EOG, and PPG (pulse signal). Many recordings were excluded from the study because they lacked one or more of these modalities, or because the available recordings exhibited insufficient signal quality, characterized by discontinuities, extensive artifacts, or corrupted segments, that would have compromised the reliability of the subsequent multimodal feature extraction and analysis. The use of 31 subjects is considered sufficient for a proof-of-concept evaluation in the early stage of model development.

The study involved a preventive signal quality assessment before feature extraction, excluding temporal segments affected by artifacts, excessive noise, or saturation to prevent bias in feature ranking and degradation of model performance. Artifact rejection was applied uniformly across EEG, EOG, and PPG signals, retaining only artifact-free segments for subsequent processing. Also, all epochs marked in the annotation file with a "−2" or "8" label were discarded. Each signal was then preprocessed using a custom MATLAB script. For the EEG signal, the MATLAB script first applied a band-pass filter (0.2–44 Hz, 4-order) to remove unwanted noise and baseline wander, then a 50 Hz notch filter to eliminate power-line interference. The script then selected the F4-F3 PSG derivation and segmented the signal into non-overlapping 5 s epochs. Each 5 s epoch was assigned the sleep stage corresponding to the original 30 s epoch in the dataset, and EEG features were extracted from each epoch for subsequent analysis.

Table 3. EEG Features extracted in the three domains, along with their corresponding physiological meanings.

Domain	Extracted Feature	Physiological Meaning
Time-domain features	Maximum Value	Statistical features are useful for analyzing the nonlinear and complex characteristics of EEGs during sleep, for discriminating between sleep stages, and for identifying differences across statistical models [39]. For instance, the standard deviation distribution shows variations in peak value, mean value, and distribution width across different sleep stages (the mean value increases from wakefulness to REM sleep up to N3). Moreover, transitions between stages are highlighted by changes in the slope of the variance.
	Minimum Value	
	Mean Value	
	Median	
	Root Mean Square	
	25th, 50th, 75th Percentile	
	Variance	
	Skewness [1,16]	The distributions of skewness and kurtosis are characteristic of each sleep stage; the skewness tails and kurtosis distributions decrease from wakefulness to REM sleep and then to N3; the peak values mark the transitions between sleep stages [39].
	Kurtosis [16]	
	Hjorth Parameters [16]	They enable the discrimination of sleep stages and the assessment of sleep quality and stability. Specifically, signal activity in the sensorimotor area differentiates wakefulness/REM (high value) from NREM (low value). In contrast, in the visual area, it distinguishes wakefulness from REM (higher activity and lower complexity). Moreover, the activity and complexity of signals in the visual cortex enable the differentiation between wakefulness and REM sleep (higher activity and lower complexity) [40]. These parameters can also be used to identify segments with unstable or alternating amplitude during NREM, whose rate indicates sleep quality: a higher rate corresponds to lower sleep quality [40].
	Zero-Crossing Rate (ZCR) [20]	The ZCR reflects sleep depth, with lower values in stages featuring slow waves (NREM) and higher values during wakefulness. Additionally, a reduction in ZCR is observed in posterior brain regions during sleep, due to a decrease in beta-band power in favor of alpha and theta bands, which are more prominent in posterior areas than in anterior ones [41].
	Average Amplitude Change (AAC) [42]	They are indices of signal complexity, with higher values during wakefulness and REM sleep and lower values during deep sleep (NREM) [43].
	Clearance Factor [42]	
	Interquartile Range (IQR) [42]	The IQR is lower during deep and REM sleep, indicating a more stable EEG, while it is higher during wakefulness and light sleep, indicating a more irregular EEG [44].
	Simple Square Integral (SSI) [42]	The SSI parameter measures the signal energy in specific sub-bands. Energy values tend to decrease from wakefulness to REM sleep, along with the signal's entropy and standard deviation [45].

Table 3. *Cont.*

Domain	Extracted Feature	Physiological Meaning
Frequency-domain features	Total Power	During deep sleep (NREM), the total power of the EEG signal is greater than in lighter sleep stages (REM). A power reduction in the delta, theta, and alpha bands indicates a state of sleep deprivation [46,47]. In subjects with significant sleep deprivation, variations in power across individual waves change with the recovery sleep cycle. The slow waves' power is greater in the second cycle than in the first, while the alpha waves' power is higher in the first cycle than in the second [48].
	Relative Spectral Power (RSP)	It allows for distinguishing between subjects with subjective and objective insomnia [49]. A significant increase in relative power in the alpha band, particularly in the right hemisphere (Fp2), indicates greater brain activation in response to stress [50].
	Power Ratios	Power ratios vary across sleep stages: higher during wakefulness and N1, lower in N2-N3, higher again in REM than in N3 [51]
	Dominant Frequency	The dominant frequency varies depending on sleep stage [52]: REM sleep and N1: the alpha band is predominant. N2 and slow-wave sleep (N3-N4): the alpha band is present, but patterns are more variable with mixed-frequency components. Slow-wave sleep: A reduction in delta power is observed compared to the baseline state and the pre-arousal interval.
	Slow Wave Index (SWI) [16]	Slow wave indices (SWIs) indicate how delta, theta, and alpha bands behave in relation to slow waves. During the transition from wakefulness to sleep, the alpha slow wave index (ASI) tends to decrease as alpha wave power declines, while the power of slow waves, such as delta and theta, increases [53,54].
	Harmonic Parameters [54]	Wakefulness: High center frequency (20 Hz), relatively wide bandwidth (15 Hz), reduced value at the center frequency. N1: Lower center frequency (15 Hz) and bandwidth (5–10 Hz) compared to wakefulness, increased presence of spindles. N2: Further reduction in center frequency (5 Hz) and bandwidth (5 Hz) with an increase in value at the center frequency. N3: Low center frequency (around 0 Hz), unchanged bandwidth compared to N2 (5 Hz), with a high value at center frequency (150–400 V^2/Hz) [55].
	Band Energy	EEG frequency bands exhibit energy variations that reflect changes in sleep stages [51]: alpha-band energy is highest during wakefulness and lowest in N3; beta band is dominant during wakefulness and lowest in N3; theta band is highest during REM sleep and lowest during N3; delta-band energy is highest in N3, lowest during wakefulness. An increase in relative power in the theta and delta bands and a decrease in Absolute Power in the alpha and beta bands indicate sleep deprivation [56].
	Spectral Slope	Spectral slope is steeper in REM sleep, indicating increased cortical inhibition, while a flattened slope is a marker of aging [57].

Table 3. *Cont.*

Domain	Extracted Feature	Physiological Meaning
	Spectral Edge Frequency (SEFd)	SEFd (mean difference between SEF95 and SEF50 across 2 s sub-epochs within a 30 s window), a novel frequency-domain feature for REM detection, quantifies the spectral spread in the 8–16 Hz band; consistent peaks in SEFd were observed during REM stages, while N2 and N3 exhibited lower values and N1 showed slightly elevated but variable trends [24].
	Absolute Power (AP)	Absolute Power (AP), calculated in the 8–16 Hz frequency range using the logarithmic sum of Fourier magnitudes, showed the lowest values during REM sleep and higher values in the Wake and N1 stages, making it a useful frequency-domain feature for distinguishing these sleep stages [24].
Non-linear domain features	Spectral Entropy [58]	These features provide additional information useful for distinguishing between psychological states and better identifying pathological conditions in sleep [59,60]. Under stress, Lempel–Ziv Complexity (LZC) is significantly higher at the Fp1 and Fp2 sites, indicating increased brain activity complexity; in contrast, Rényi entropy is lower, suggesting a reduction in the global complexity of brain dynamics, indicative of a more rigid brain system under stress [50].
	Singular Value Decomposition (SVD) Entropy [61]	
	Lempel–Ziv Complexity [62]	
	Rényi Entropy [58]	

Table 4. EOG Features extracted in the two domains, along with their corresponding physiological meanings.

Domain	Extracted Feature	Physiological Meaning
Time-domain features	Blink Duration: Mean Value [10]	Blink frequency and blink duration show characteristic changes with increasing sleepiness. In particular, as fatigue increases, blink frequency tends to decrease, while overall blink duration increases significantly [10].
	Blink Duration: Variance [10]	
	Blink Duration: Kurtosis [10]	
	Blink duration: Standard deviation [10]	
	Closing time [10]	
	Reopening time [10]	
	Blink frequency [10]	

Table 4. *Cont.*

Domain	Extracted Feature	Physiological Meaning
Frequency-domain features	Slow Eye Movements (SEMs) [11]	SEM reflect the transition from wakefulness to sleep, with reduced cortical control of eye movements. Their predominance in NREM1 indicates a transitional stage closer to wakefulness, whereas their reduction in NREM2–3 reflects stable, deeper sleep.
	Power at 0.5 Hz [63]	EOG power in the lowest delta range, reflecting extremely slow eye movements typical of sleep onset [63].
	Power at 5 Hz [63]	Theta-band EOG activity associated with the wake-sleep transition and the emergence of SEM [63].
	Power at DeltaTheta (0.5–7 Hz)	Overall, slow EOG activity increases with sleepiness and cortical synchronization [63].
	Power at BetaGamma (13–40 Hz)	Fast EOG components linked to cortical activation and vigilance, decreasing with increasing sleepiness [63].
	Ratio DeltaTheta/BetaGamma	Ratio between slow and fast EOG activity; higher values indicate increased sleepiness and reduced vigilance [63].

Table 5. Time domain features extracted from the respiratory component of the PPG signal, along with their corresponding physiological meanings.

Domain	Extracted Feature	Physiological Meaning
Time-domain features	Breathing Rate (BR) [12]	Describe respiratory rhythm and stability across sleep stages: slower and more regular in NREM, faster and more irregular in REM and wakefulness; the inspiratory/expiratory ratio reflects respiratory pattern variability [12–14].
	Inspiratory-to-Expiratory Ratio (IER) [13]	

Table 6. Time- and frequency-domain features extracted from the cardiac component of the PPG signal, along with their corresponding physiological meanings.

Domain	Extracted Feature	Physiological Meaning
Time-domain features	Heart Rate (HR) [15]	Heart rate decreases progressively during NREM sleep due to increased parasympathetic (vagal) control, and increases during REM sleep, reflecting increased sympathetic activity [15,64].
	Standard Dev. of Consecutive PP Interval Diff. (SD1) [15]	SD1 quantifies short-term variability of consecutive PP intervals and primarily reflects parasympathetic activity, which is dominant in NREM sleep and reduced during REM sleep [15,64].
	Standard Deviation of NN Intervals (SDNN) [15]	SDNN represents overall heart rate variability, capturing both sympathetic and parasympathetic influences, and is lowest during slow-wave sleep due to strong vagal modulation [15,64].
	SDRatio: SD1/SD2 (SD2 = Long-Term Poincaré Variability) [15]	The ratio between short- and long-term HRV components indicates the balance between fast and slow heart rate fluctuations and reflects autonomic regulation across different sleep stages [45,46].
	Root Mean Square of Successive Differences (RMSSD) [15]	RMSSD measures short-term PP interval variations and serves as an index of parasympathetic (vagal) activity, which predominates during NREM sleep [45,46].
Frequency-domain features	Low-Frequency (0.04–0.15 Hz) [15]	LF power reflects both sympathetic and parasympathetic modulation of heart rate and contributes to the LF/HF ratio as a marker of autonomic balance across sleep stages [45,46].
	High-Frequency (0.15–0.4Hz) [15]	HF power is associated with parasympathetic (vagal) activity, particularly pronounced during slow-wave NREM sleep, and decreases during REM sleep [45,46].
	LF-HF Ratio [15]	The LF/HF ratio reflects the balance between sympathetic and parasympathetic activity: it is low during NREM sleep due to parasympathetic predominance and high during REM sleep, when sympathetic activity increases [45,46].

The EOG signal was preprocessed to extract blink-related features. First, the signal was filtered (EOG_{filt}) using a band-pass filter (0.5–10 Hz, 8-order) and then normalized (EOG_{norm}) as follows:

$$EOG_{norm} = \frac{EOG_{filt} - \mu}{\sigma} \quad (1)$$

where μ and σ are the mean and standard deviation of the filtered signal, respectively, computed over each 5 s analysis window. Using the normalized signal as a starting point, a dynamic threshold was defined as

$$Threshold = k(\mu + a \times \sigma) \quad (2)$$

where a = 2 and k = 0.25 are scale factors, empirically selected through validation using manually annotated EOG signals. The threshold was updated every 5 s to adapt to local signal characteristics. The samples that exceed this threshold have been identified and grouped into supra-threshold regions, which may represent blinking events. In each location, the blink peak is determined by searching for the local maximum, from which the blink's amplitude is calculated. A blink was considered valid if three criteria were met: the amplitude of the peak had to exceed a predetermined minimum, there had to be a negative slope following the peak, and the total duration of the event had to be equal to or greater than 100 ms. After verification, the algorithm identified the eye reopening time as the local minimum within 500 ms of the event offset, corresponding to the onset of the reopening phase. It also determined the closure time by identifying the signal's zero-crossing before blink initiation, marking the onset of the closure phase. Figure 1 shows a representative example of the blink validation process, including peak detection, zero-crossing identification, and estimation of reopening and closing times. Finally, the algorithm segmented the signal into non-overlapping 5 s epochs, assigned each epoch the sleep stage corresponding to the original 30 s epoch in the dataset, and extracted the EOG features listed in Table 4 for subsequent analysis.

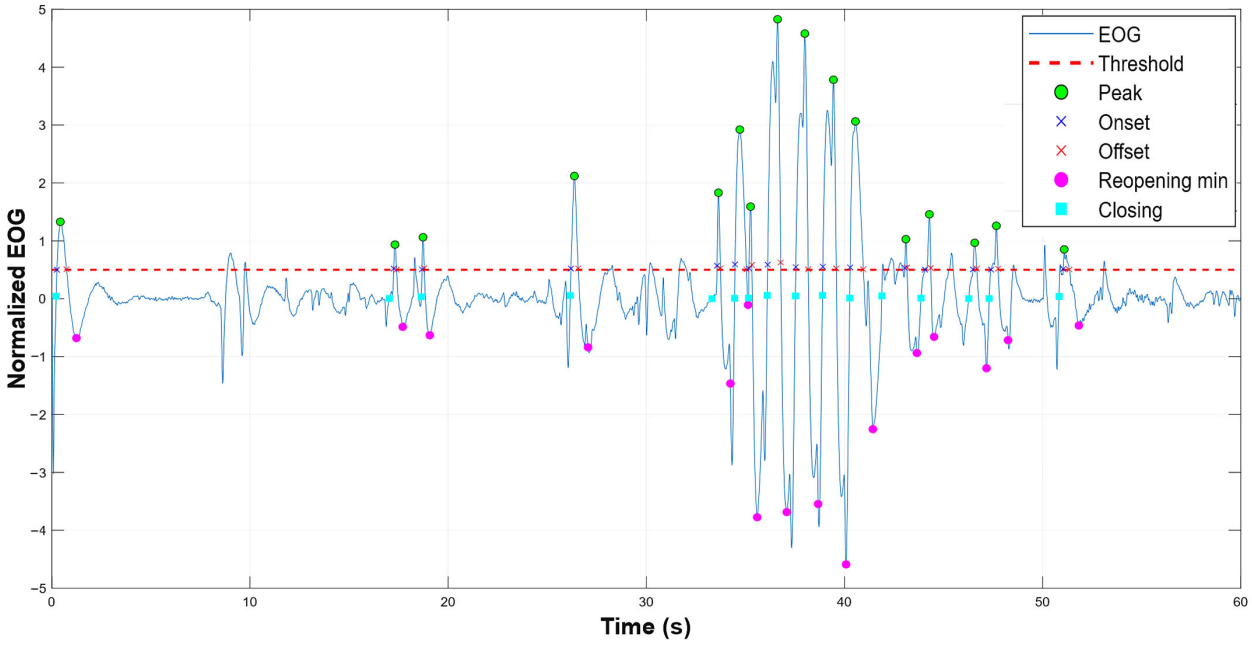


Figure 1. Identification of EOG blinks, where the dotted line indicates the threshold of 0.5 (in this example); the blue and red "×" mark the start and end times of the blink; finally, the closing and reopening times are indicated by the light blue squares (closing time) and the fuchsia circles (reopening time).

The analysis extracted respiratory features from the PPG signal by computing the respiratory envelope using the Hilbert transform and a band-pass filter at 0.1–0.5 Hz (order 8) to isolate respiration-related oscillations. The resulting signal was then temporally segmented into non-overlapping 30 s windows. Within each window, respiratory cycles were identified using an adaptive threshold set to 70% of the signal's Root Mean Square (RMS) value. This threshold value was empirically selected during the validation phase by evaluating the correlation between the detected respiratory events and the reference signal.

This approach enabled the detection of positive peaks and negative troughs of the signal, corresponding to the inspiratory and expiratory phases, respectively (as shown in Figure 2). Respiratory cycles were defined as the interval between two consecutive troughs, enabling the extraction of key respiratory features, such as the Breathing Rate (BR) and the Inspiratory-to-Expiratory Ratio (IER). These features were averaged over 30 s windows and associated with the corresponding sleep stage. To ensure temporal consistency with features extracted from EEG and EOG signals, each sample was replicated 6 times, yielding a final temporal resolution of 5 s.

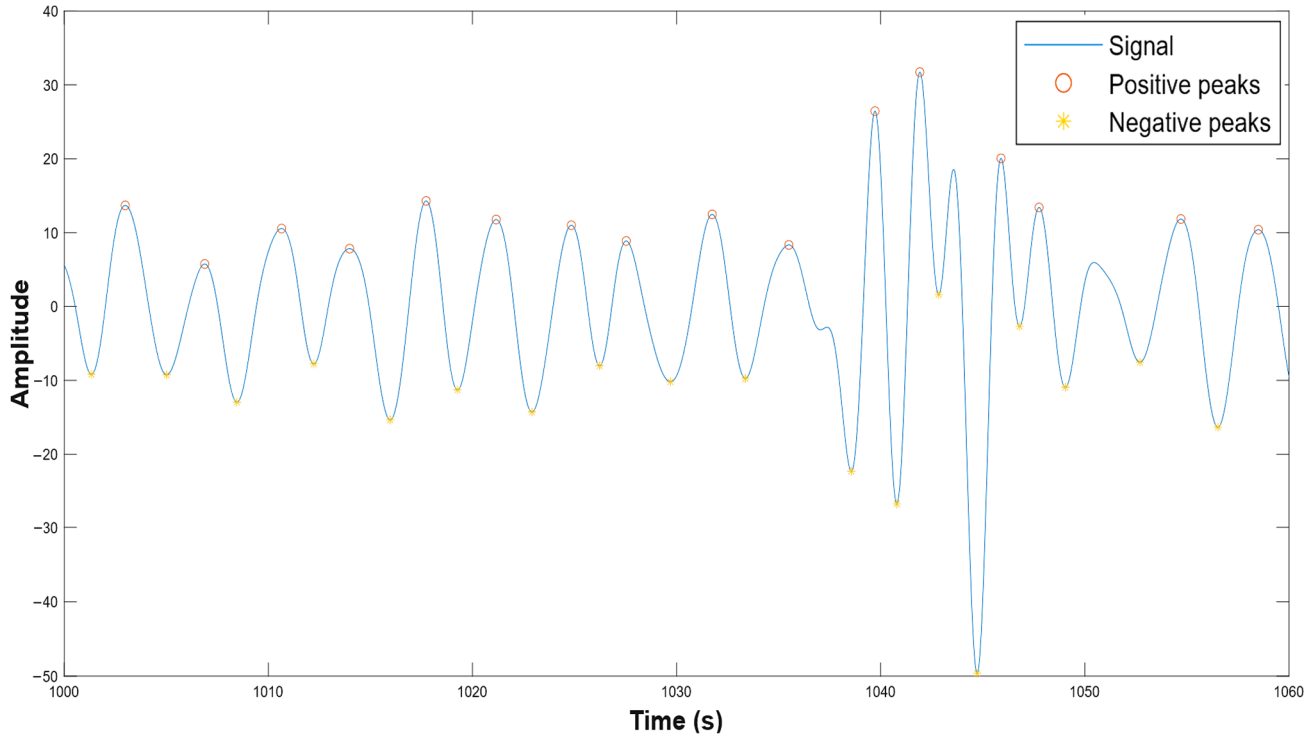


Figure 2. Positive inspiratory peaks (red circles) and expiratory peaks (yellow "*") identified in the respiratory signal extracted from PPG.

Finally, cardiac features from the PPG signal were extracted by identifying heartbeats, which correspond to the positive peak signals (indicated by circles in Figure 3). During preprocessing, the PPG signal was band-pass filtered between 0.5 and 4 Hz (8-order) to isolate the cardiac component. The filtered signal was then segmented into non-overlapping 30 s windows. Within each window, heartbeats were detected using an adaptive threshold set to 25% of the filtered signal's RMS. This threshold value was empirically determined by maximizing the correlation between the detected cardiac events and the reference signal. To ensure reliable peak detection, additional constraints were imposed, including a minimum inter-peak distance of 0.3 s and a minimum peak prominence equal to the adaptive threshold. Based on the detected peak timestamps, Pulse-to-Pulse (PP) intervals

were computed as the time difference between two consecutive heartbeats, according to Equation (3):

$$PP_n = t_{n+1} - t_n \quad (3)$$

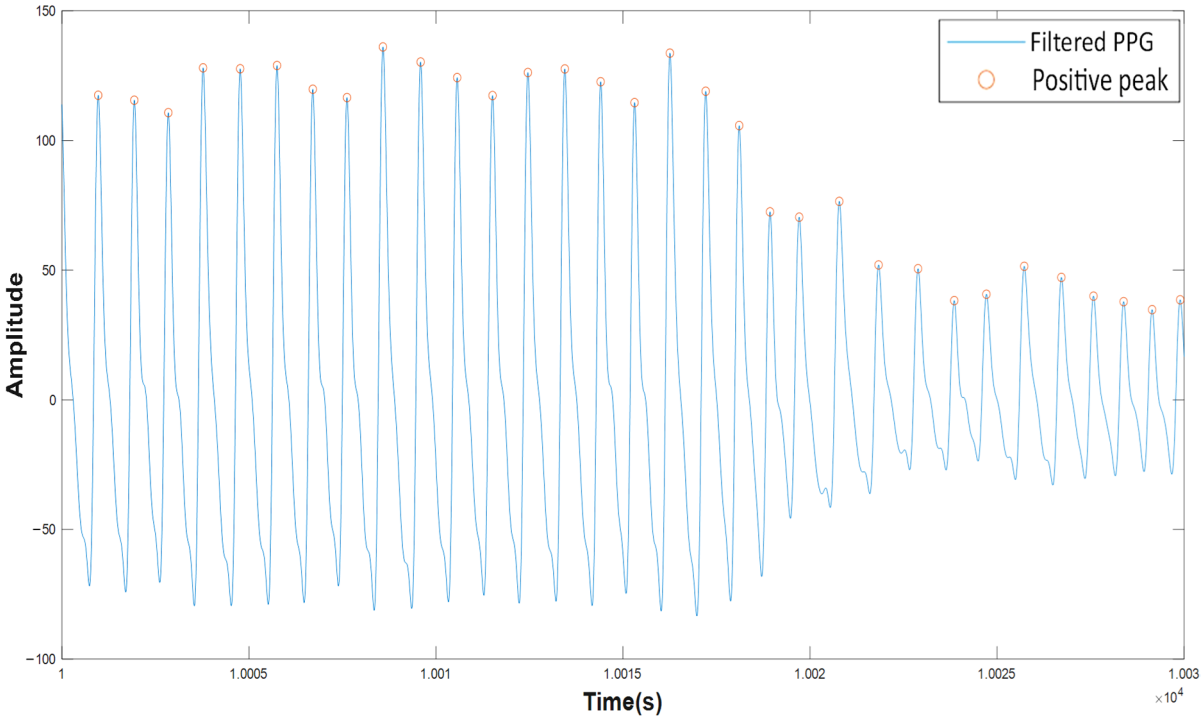


Figure 3. Cardiac peaks (red circles) detected in the PPG-derived cardiac component.

Using the resulting PP intervals, the cardiac features listed in Table 6 were extracted. These features were subsequently averaged over 30 s windows and associated with the corresponding sleep stage. Finally, to ensure temporal consistency with the other extracted features, each sample was replicated 6 times, yielding a final temporal resolution of 5 s. Figure 3 depicts the process for detecting positive peaks related to the cardiac component, which is useful for extracting the corresponding features.

3.3. Feature Optimization and Dimensionality Reduction Strategies

Subsequently, the selected features were extracted from the EEG and EOG signals using MATLAB (version r2024b), and the mRMR (Minimum Redundancy Maximum Relevance) algorithm was applied for feature selection. To ensure strict independence between model training and evaluation, the complete dataset was subject-wise split into training (24 subjects) and test (7 subjects) sets before applying any data-dependent transformations. All preprocessing steps, including standardization via StandardScaler, feature selection using mRMR, and dimensionality reduction via PCA, were applied exclusively on the training set. Specifically, mRMR feature relevance and redundancy calculations (Equations (4)–(6)), as well as the 0.01 selection threshold, were evaluated solely on training samples. Then, the PCA transformation matrix and eigenvectors (retaining 98% of the total variance) were computed using only the selected training features. The held-out test dataset was treated as unseen data: it was scaled using the mean and variance learned from the training set, filtered according to the training-derived mRMR feature indices, and projected into the principal component space using the training-derived PCA eigenvectors. This procedure ensures that no information from the test set influenced feature selection, dimensionality reduction, or any other preprocessing step.

The mRMR algorithm is a feature selection technique that identifies an optimal subset of features with higher information content than the target variable and that are minimally redundant with one another. The algorithm evaluates feature relevance using Mutual Information (MI), which quantifies the statistical dependence between a feature and the target variable, ensuring that selected features contribute meaningful, discriminative information. At the same time, mRMR accounts for redundancy by measuring the mutual information between pairs of features, thus discouraging the inclusion of highly correlated or overlapping variables.

The objective is to determine a feature subset (S) that maximizes the average relevance V_S with respect to the target variable (y) and minimizes the average redundancy W_S within the subset, defined as

$$V_S = \frac{1}{|S|}\sum_{x \in S} I(x, y) \tag{4}$$

$$W_S = \frac{1}{|S| \cdot (|S| - 1)}\sum_{x,z \in S, x \neq z} I(x, z) \tag{5}$$

where $|S|$ denotes the number of selected features, $I(x,\ y)$ represents the mutual information between the feature x and the target variable y, and $I(x,\ z)$ is the mutual information between pairs of features x and z, with z indicating another feature belonging to the same subset S. The feature selection process iteratively selects the feature with the highest relevance to the target variable, and subsequently incorporates additional features by optimizing the Mutual Information Quotient (MIQ), defined as

$$MIQ_x = \frac{V_x}{W_x} \tag{6}$$

where V_x ($I(x,y)$) and W_x ($1/|S| \sum_{z \in S} I(x,z)$) represent the relevance and redundancy of the candidate feature x, respectively. By prioritizing complementary and non-redundant features, the mRMR algorithm improves model interpretability and computational efficiency. These properties make it well-suited for high-dimensional datasets, such as those comprising EEG and EOG features [65,66]. In detail, the fscmrmr function in MATLAB's Statistics and Machine Learning Toolbox was exploited, which implements the mRMR algorithm for feature selection. The fscmrmr function accepts the feature table and the corresponding response variable (i.e., sleep stages) as arguments. This implementation computes mutual information only between distinct feature pairs ($x \neq z$), as mutual information with the feature itself would not provide a meaningful measure of inter-feature redundancy (Equation (5)). The function also implements a sequential selection algorithm that avoids division by zero. Features with zero redundancy are prioritized before the MIQ criterion is applied, and numerical safeguards are included for stability. In practice, all selected features had $W_x > 0$ in our dataset. Therefore, the function outputs the feature rank and the corresponding scores to which the thresholding was applied.

PCA reduces data complexity by transforming the original features into a new set of uncorrelated variables, called principal components (PCs), obtained as linear combinations of the original features. Each principal component corresponds to a direction in the feature space along which the data variance is maximized. The first principal component captures the largest amount of variance, while each subsequent component, orthogonal to the previous ones, explains the maximum remaining variance. Although the complete set of PCs preserves the original dimensionality, PCA is commonly used for dimensionality reduction by retaining only the components that explain a significant portion of the total variance. The cumulative explained variance typically guides this selection. By projecting the data onto a lower-dimensional subspace, PCA reduces feature redundancy, suppresses noise, and preserves the most informative structures of the data. These properties make

PCA particularly advantageous for high-dimensional datasets, improving computational efficiency, model robustness, and overall classification performance.

3.4. Architecture of the Two-Stage DL-Based Framework for Sleep Staging

The proposed algorithm for automatic sleep staging follows a two-stage DL framework based on LSTM networks with an attention mechanism. In the first stage, a WRN classifier assigns each input multimodal sequence, including PCs of EEG and EOG signals, as well as respiratory and cardiac features extracted from the PPG signal, to one of the three macro sleep stages: Wake, REM, or NREM. In the second stage, only the sequences classified as NREM are further processed by a dedicated NREM classifier to distinguish among the three NREM sub-stages (N1, N2, and N3).

Therefore, the hierarchical combination of the two models enables the classification of multimodal sequences into the five canonical sleep stages: Wake, REM, N1, N2, and N3. Figure 4 presents the overall classification pipeline: each sequence is first processed by the WRN model, and only those classified as NREM are passed to the second model for fine-grained classification (N1, N2, N3). This modular design reduces the complexity of the five-class classification problem by decomposing it into two simpler tasks, improving class balance (as described in Section 4) and allowing each model to specialize, ultimately enhancing the overall performance of the system.

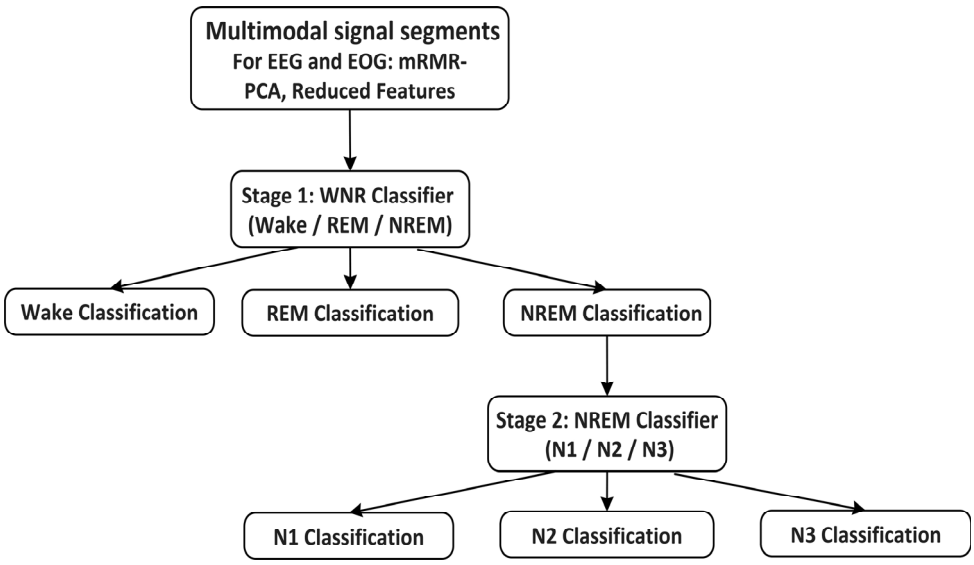


Figure 4. Block diagram of the combined 5-class sleep staging model.

Both classifiers share a similar architecture but differ in their configurations. The WRN model employs two stacked LSTM layers with 128 and 64 units, respectively, while the NREM model uses 96 and 48 units. Each LSTM layer is followed by batch normalization and dropout layers to improve training stability and reduce overfitting. The output of the second LSTM layer is fed into an attention mechanism (highlighted in Figure 5), implemented through a fully connected layer with tanh activation, followed by a softmax operation applied along the temporal dimension. The resulting attention weights are reshaped using repeat and permute operations and combined with the LSTM outputs via element-wise multiplication. This mechanism enables the model to emphasize the most relevant temporal features while attenuating less informative ones, improving the detection of critical patterns in multimodal signals. Following the attention module, the extracted features are aggregated via a global average pooling layer and further processed by two fully connected layers with 64 and 32 neurons, respectively, both activated by ReLU

(Rectified Linear Unit), followed by batch normalization and dropout layers. The final output is produced by a softmax layer, with three output classes for both the WRN (Wake, REM, and NREM) and NREM (N1, N2, and N3) classifiers. Figure 5a illustrates the architecture of the WRN classifier, while Figure 5b shows the NREM classifier.

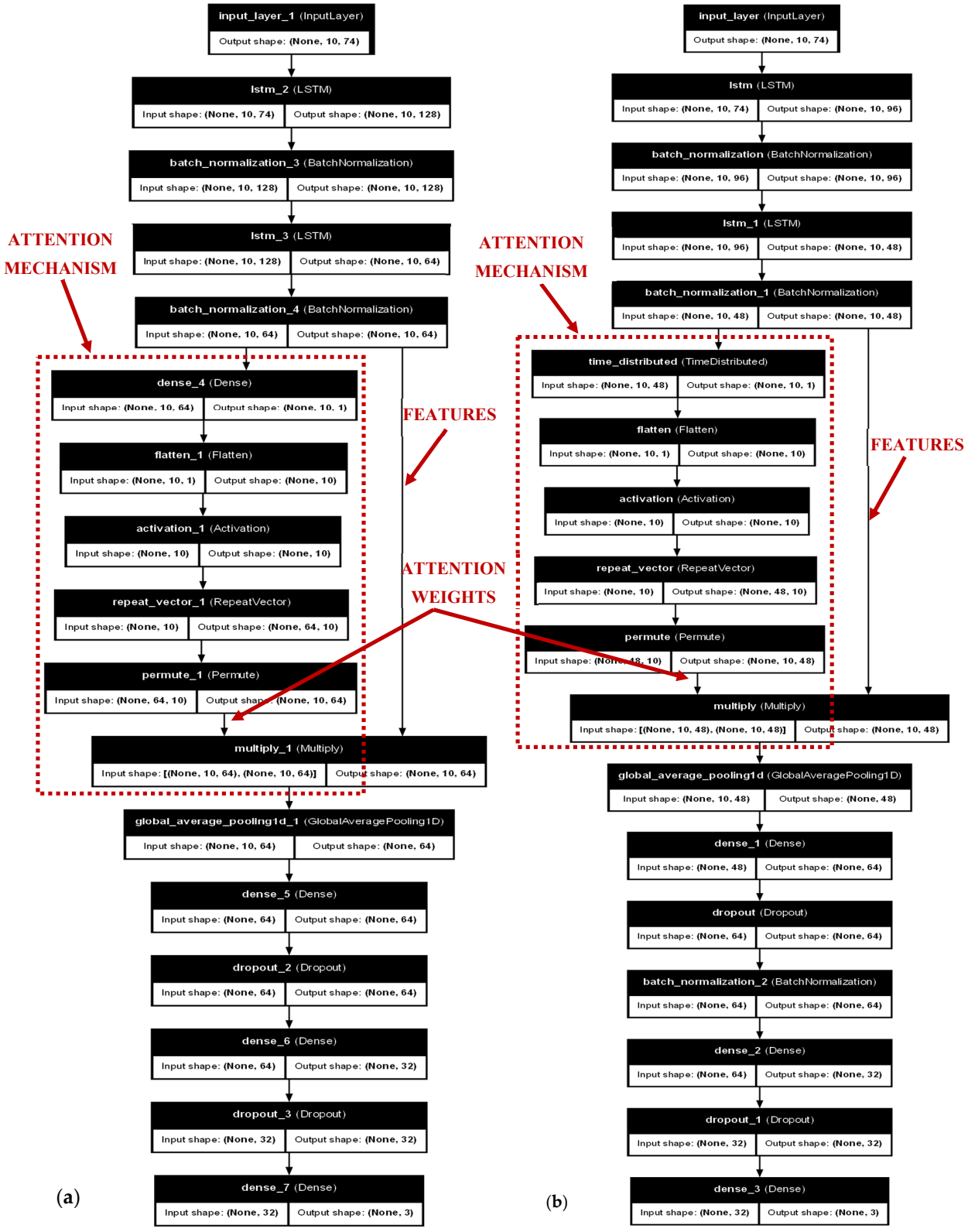


Figure 5. (**a**) The architecture of the model for the WRN classifier, and (**b**) for the NREM classifier.

3.5. Training and Testing Dataset Construction

Before sequence construction, the selected 31 subjects from the BOAS dataset were partitioned to ensure complete independence between training and testing data. The training set comprised recordings from 24 subjects, while the independent test set consisted of 7 independent subjects (IDs: 72, 62, 82, 88, 71, 87, 49). As previously discussed, all data-dependent processing steps, including mRMR-based feature selection, dimensionality reduction via PCA, and feature standardization, were performed exclusively using the training subjects. The learned transformations were then applied unchanged to the independent test set to evaluate model performance.

The training set comprised 126,607 labeled 5 s sub-epochs, distributed as follows: 12,795 Wake, 25,286 REM, and 88,526 NREM (Table 7, Step 1). Within the NREM class, 2845 sub-epochs belonged to N1, 80,155 to N2, and 5526 to N3 (Table 8, Step 1). Figure 6 illustrates the class distribution of the training data, highlighting the marked imbalance between Wake, REM, and NREM, as well as the under-representation of the N1 and N3 stages relative to N2. Tables 7 and 8 summarize the composition of the datasets used for training the WRN and NREM classifiers after sequence generation and data augmentation. For both three-stage classifiers (WRN and NREM), to capture temporal dependencies of features from sleep stages, the epochs were aggregated into fixed-length sequences. Sequences were constructed using a sliding-window approach, grouping ten consecutive 5 s epochs into 50 s windows.

Table 7. Composition and distribution of the training set for the WRN model, following the data augmentation and sliding window steps.

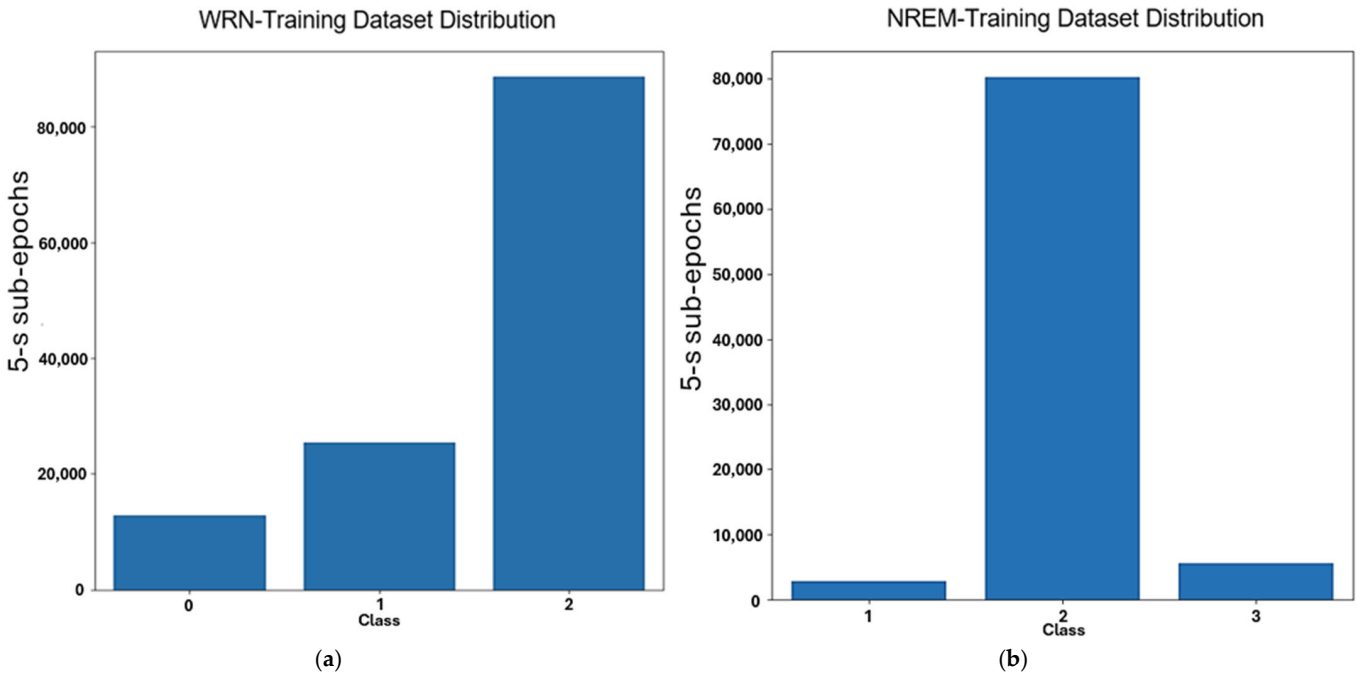
Step	Description	Total Samples/Sequences	Class Wake	Class REM	Class NREM
1	Original dataset	126,607 samples	12,795 samples	25,286 samples	88,526 samples
2	After sliding window (timesteps = 10, stride = 5)	20,256 sequences	2034 sequences	4049 sequences	14,173 sequences
3	Augmentation	56,738 sequences	10,170 sequences	4049 sequences	42,519 sequences

Table 8. Composition and distribution of the training set for the NREM model, following the data augmentation and sliding window steps.

Step	Description	Total Samples/Sequences	Class N1	Class N2	Class N3
1	Original dataset	88,526 samples	2845 samples	80,155 samples	5526 samples
2	After sliding window (timesteps = 10, stride = 10)	7081 sequences	226 sequences	6420 sequences	435 sequences
3	Augmentation	8324 sequences	1356 sequences	6420 sequences	1740 sequences

For the WRN classifier, consecutive sequences were constructed using a sliding-window approach with 10 timesteps (50 s) and stride = 5, allowing partial overlap between neighboring sequences only within the same dataset partition. Because the subject-wise split was performed before sequence generation, no overlapping sequences or feature vectors were shared between the training and test sets, thereby eliminating the possibility of information leakage. For the NREM classifier, non-overlapping sequences were generated by directly grouping consecutive feature vectors into fixed-length windows (timesteps = 10, stride = 10). This strategy was adopted to preserve the independence of neighboring sequences while providing stable discrimination among the more subtle NREM sub-stages.

The subject-wise partitioning, together with the independent preprocessing pipeline, ensures that the reported performance reflects the model's ability to generalize to previously unseen subjects.



(**a**) (**b**)

Figure 6. Class distribution of the 5 s sub-epochs used to build the datasets for training the WRN and NREM classifiers, highlighting the imbalance among sleep stages. (**a**) Shows the distribution of Wake (class 0), REM (class 1), and NREM (class 2) epochs; (**b**) reports the distribution of NREM sub-stages, where class 1 corresponds to N1, class 2 to N2, and class 3 to N3.

Five-second epochs were adopted to provide finer temporal sampling of the physiological signals and to construct longer temporal sequences for LSTM modeling. Since the BOAS dataset provides labels at a 30 s resolution, each 5 s sub-epoch inherited the label of its parent 30 s epoch as scored by expert consensus; therefore, the ground-truth annotation resolution remained 30 s, and within-epoch stage transitions could not be explicitly resolved. Consequently, the 5 s segmentation was used solely to provide finer temporal sampling at the feature level for sequence-based modeling. This choice is consistent with prior DL-based sleep staging studies that employ sub-epoch segmentation for feature extraction and temporal sequence modeling [67–69]. Therefore, the 5 s sub-epochs are not statistically independent observations, since each group of six consecutive sub-epochs inherits the label of the same expert-scored 30 s epoch. Overall, the training dataset comprises 21,101 unique 30 s reference epochs, from which the 5 s sub-epochs and the corresponding 50 s input sequences were generated.

For the WRN classifier, a 5 sub-epoch stride was used during sequence construction, resulting in partially overlapping windows (a 50%/25 s overlap). In contrast, for the NREM classifier, a non-overlapping approach (stride = 10) was used. Each sequence was labeled using majority voting over its 10 constituent 5-s sub-epochs (t_1 to t_{10}), assigning each 50 s-window the most frequent class across the ten sub-epochs. In the case of a 5:5 tie (e.g., five sub-epochs of class A and five of class B), the tie was deterministically resolved by considering the second half of the sequence window (sub-epochs t_6 to t_{10}). Specifically, the class appearing most frequently in the latter half of the window was selected as the sequence label. This tie-breaking rule was chosen to reflect the temporal dynamics of sleep stage transitions, as the latter portion of the window is temporally closer to the midpoint

of the sequence and better represents the ongoing stage. This deterministic tie-breaking is applied consistently across all sequences, ensuring reproducibility and avoiding any randomness in the labeling process. The assigned label represents the majority class over the 50 s window, with the temporal midpoint at t = 25 s.

All training and testing are performed at the sequence level (50 s windows). Each 50 s sequence is treated as a simple sample for both training and testing. Therefore, the performance metrics reported in Section 4.2 correspond to sequence-level classification, where the support values in confusion matrices refer to the number of 50 s sequences evaluated. This sequence-level evaluation is consistent with prior DL sleep staging studies [67–69]. The 50 s sequence length provides sufficient temporal context for the LSTM to capture sleep transition patterns while maintaining a manageable sequence length for recurrent architecture.

To address class imbalance, oversampling was applied exclusively to the training sets using data augmentation techniques, including Gaussian noise, feature scaling, temporal shifts, masking of temporal segments, and amplitude attenuation of randomly selected features (reducing their amplitude by 30–70%), combined with a class-weighted loss function (Balanced Crossentropy) computed from the original class distribution. The parameters of the implemented augmentation strategy are summarized in Table 9. The augmentation operations were applied in the following fixed order: Gaussian noise → feature scaling → circular shift → temporal masking → channel dropout.

Table 9. Augmentation and oversampling parameters for each sleep stage class, including noise levels, scaling ranges, probabilities, and temporal augmentation settings. The oversampling multiplier indicates the factor by which each class was synthetically augmented to address class imbalance. REM and N2 (multiplier = 1×) were not oversampled but still received online augmentation (noise and scaling) as regularization. "—" indicates that the augmentation was not applied.

Parameter	WRN-Wake	WRN-NREM	WRN-REM	NREM-N1	NREM-N3	NREM-N2
Noise Level	0.08	0.04	0.02	0.10	0.08	0.03
Scale Range	(0.75, 1.25)	(0.85, 1.15)	(0.95, 1.05)	(0.70, 1.30)	(0.75, 1.25)	(0.90, 1.10)
Shift Probability	0.80	0.60	0.40	0.90	0.80	0.30
Shift Range	[−4, +4] timesteps	[−4, +4] timesteps	[−4, +4] timesteps	[−4, +4] timesteps	[−4, +4] timesteps	[−4, +4] timesteps
Mask Probability	0.50	0.30	0.20	0.70	0.50	0.10
Mask Length	[2, 5) timesteps	[2, 5) timesteps	[2, 5) timesteps	[2, 5) timesteps	[2, 5) timesteps	[2, 5) timesteps
Channel Dropout Prob.	0.60	0.60	—	0.60	0.60	—
Channel Dropout Range	[1, min(4, n_channels)]	[1, min(4, n_channels)]	—	[1, min(4, n_channels)]	[1, min(4, n_channels)]	—
Attenuation Range	(0.30, 0.70)	(0.30, 0.70)	—	(0.30, 0.70)	(0.30, 0.70)	—
Oversampling Multiplier	5×	3×	1×	6×	4×	1×

For the WRN classifier, Wake (2034 sequences, 10.1% of training data) was augmented by a factor of 5, and NREM (14,173 sequences, 69.9%) by a factor of 3. Although NREM is the majority class at the macro-level, its augmentation was necessary because it serves as the gateway to the NREM sub-stages (N1, N2, and N3). By improving the WRN classifier's ability to identify NREM patterns correctly, more NREM epochs (particularly the rare N1

(3.2%) and N3 (6.1%) stages reach the NREM classifier for refinement. REM (4049 sequences, 20.0%) was not augmented, as it was already sufficiently represented. This hierarchical balancing strategy aligns with the class imbalance literature, which recommends addressing imbalance at all levels of hierarchical classification [70,71].

For the NREM classifier, N1 (226 sequences, 3.2%) was augmented by a factor of 6 and N3 (435 sequences, 6.1%) by a factor of 4, while N2 (6420 sequences, 90.7%) remained unchanged. The selected augmentation factors were determined through an iterative optimization process aimed at maximizing the total accuracy of the cascaded 5-class classifier. The N1 and N3 classes were also assigned higher weights in the loss function to further increase the contribution of errors on these minority subclasses during optimization. Feature normalization was performed using a StandardScaler fitted only on the original training data before augmentation.

As described above, to assess the performance of the developed sleep stage classifiers, test sets were created from data from the 7 independent subjects from the BOAS database. The seven recordings comprised 39,029 5 s sub-epochs, from which 7804 temporal sequences were generated using the adopted sliding-window strategy. The structure of the test sets used to evaluate the WRN and NREM classifiers before and after sliding window operations is shown in Table 10.

Table 10. Composition of the test sets derived from 7 unseen subjects of the BOAS database to assess performance of WRN and NREM classifiers.

Test Set WRN	**Wake**	**REM**	**NREM**	**Total**
Original dataset	6366 samples	6968 samples	25,695 samples	39,029 samples
After sliding window (timesteps = 10, stride = 5)	1276 sequences	1397 sequences	5131 sequences	7804 sequences
Test Set NREM	**N1**	**N2**	**N3**	**Total**
Original dataset	829 samples	23,864 samples	1002 samples	25,695 samples
After sliding window (timesteps = 10, stride = 10)	79 sequences	2387 sequences	103 sequences	2569 sequences

The performance of the 5-class hierarchical classifier was evaluated using a cascade evaluation procedure. First, the WRN classifier was applied to all 7804 sequences of the WRN test set. Samples predicted as Wake or REM were directly assigned to their corresponding final class. Samples predicted as NREM were subsequently passed to the NREM classifier, which assigned the final prediction as N1, N2, or N3. Consequently, the second-stage classifier operated exclusively on samples predicted as NREM by the WRN classifier, allowing error propagation from the first stage to be fully captured in the final five-class evaluation. The resulting aligned test set consisted of 5242 sequences (Table 11). Algorithm 1 summarizes this complete cascade evaluation procedure.

The distribution of the independent test set for the 5-class sleep classifier, consisting of 50 s sequences generated from seven unseen BOAS subjects, is reported in Table 11.

The Python (ver. 3.14) script for training and testing the two-stage sleep staging model is available in the Supplementary Material (File S1).

Algorithm 1: Five-Class Cascade Evaluation

Input:

X_wrn_test: WRN test features (7804 sequences × 56 PCs × 10 timesteps)
y_wrn_test_gt: WRN ground truth labels (0 = Wake, 1 = REM, 2 = NREM)
X_nrem_test: NREM test features (2569 sequences × 56 PCs × 10 timesteps)
y_nrem_test_gt: NREM ground truth labels (0 = N1, 1 = N2, 2 = N3)
wrn_model: Trained WRN classifier (3 classes: Wake/REM/NREM)
nrem_model: Trained NREM classifier (3 classes: N1/N2/N3)
mapping_dict: Dict mapping WRN test indices → NREM test indices
• Creation: Match by (subject_id, epoch_id)
• Count: 1328 valid mappings (only NREM samples with test counterparts)

Output:

y_true_5class: Combined ground truth (5 classes: 0 = Wake, 1 = REM, 2 = N1, 3 = N2, 4 = N3)
y_pred_5class: Combined predictions (5 classes)
accuracy_5class: Overall five-class accuracy
cm_5class: 5 × 5 confusion matrix

```
begin
    P_wrn = wrn_model.Predict(X_wrn_test);
    ŷ_wrn = ArgMax(P_wrn, axis = 1);
    for each sequence index i in ŷ_wrn do
        if i not in mapping_dict then
            continue;
        end
        j = mapping_dict[i];
        if y_wrn_test_gt[i] == 0 then y_true = 0;  // Wake
        else if y_wrn_test_gt[i] == 1 then y_true = 1;  // REM
        else y_true = y_nrem_test_gt[j] + 2;  // Map NREM (0- > 2: N1, 1- > 3: N2, 2- > 4: N3)
        Append y_true to y_true_5class;
        if ŷ_wrn[i] == 0 then y_pred = 0;  // Predicted Wake
        else if ŷ_wrn[i] == 1 then y_pred = 1;  // Predicted REM
        else
            P_nrem = nrem_model.Predict(X_nrem_test[j]);
            ŷ_nrem = ArgMax(P_nrem);
            y_pred = ŷ_nrem + 2;  // Map NREM sub-class prediction to 5-class index
        end
        Append y_pred to y_pred_5class;
    end
    accuracy_5class = ComputeAccuracy(y_true_5class, y_pred_5class);
    cm_5class = ComputeConfusionMatrix(y_true_5class, y_pred_5class);
    return y_true_5class, y_pred_5class, accuracy_5class, cm_5class;
end
```

Table 11. Distribution of sleep stages in the independent subject-wise test set after cascade evaluation. The table reports the number and percentage of 50 s sequences for each of the five sleep stages. The test set comprises 5242 sequences derived from seven independent BOAS subjects.

Sleep Stage	Total Sequences	Percentage (%)
Wake	1276	24.3
REM	1397	26.7
N1	79	1.5
N2	2387	45.5
N3	103	2.0
Total	5242	100.0

3.6. Performance Evaluation Metrics

The performance of the proposed sleep staging framework was evaluated using several standard classification metrics, including total accuracy, precision, recall, F1-score, and Cohen's κ coefficient. These metrics provide complementary insights into classification performance, particularly in the presence of class imbalance across sleep stages. Total accuracy is computed as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. For each sleep stage, precision and recall are calculated as

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

The F1-score, defined as the harmonic mean of precision and recall, is computed as

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

In addition to these metrics, Cohen's κ coefficient was used to quantify the agreement between the predicted and reference sleep-stage labels, while accounting for chance agreement. Cohen's κ is defined as

$$\kappa = \frac{Pr(o) - Pr(e)}{1 - Pr(e)} \tag{11}$$

where $Pr(o)$ is the observed agreement between the predicted and true labels, and $Pr(e)$ is the expected agreement by chance. A κ value of 1 indicates perfect agreement, whereas a value of 0 corresponds to agreement equivalent to random chance.

The combination of these evaluation metrics provides a comprehensive assessment of the proposed framework by considering overall classification performance, class-wise discrimination capability, and prediction reliability beyond chance agreement.

4. Results

This section describes the selection and optimization of features extracted from EEG and EOG signals for sleep staging. A combination of mRMR and PCA was used to identify informative, non-redundant features, reduce dimensionality, and retain relevant

information for the Deep Learning model. Both EEG and EOG feature sets underwent this procedure, yielding an optimized multimodal input for 5-class sleep-stage classification.

4.1. Selection and Optimization of EEG and EOG Features

The mRMR algorithm was then applied exclusively to the EEG features of the training set, assigning each feature a score based on its redundancy and relevance, thereby identifying the most relevant and least redundant features (Figure 7). Then, applying a 0.01 threshold reduced the number of EEG features from 100 to 90, retaining only those with the highest informative contribution (Table 12). After selecting features with higher scores using the mRMR algorithm, PCA was applied to extract the PCs, which are linear combinations of the input variables (i.e., the features) and mutually orthogonal, resulting in an optimized reduced feature set for training and testing the DL algorithm.

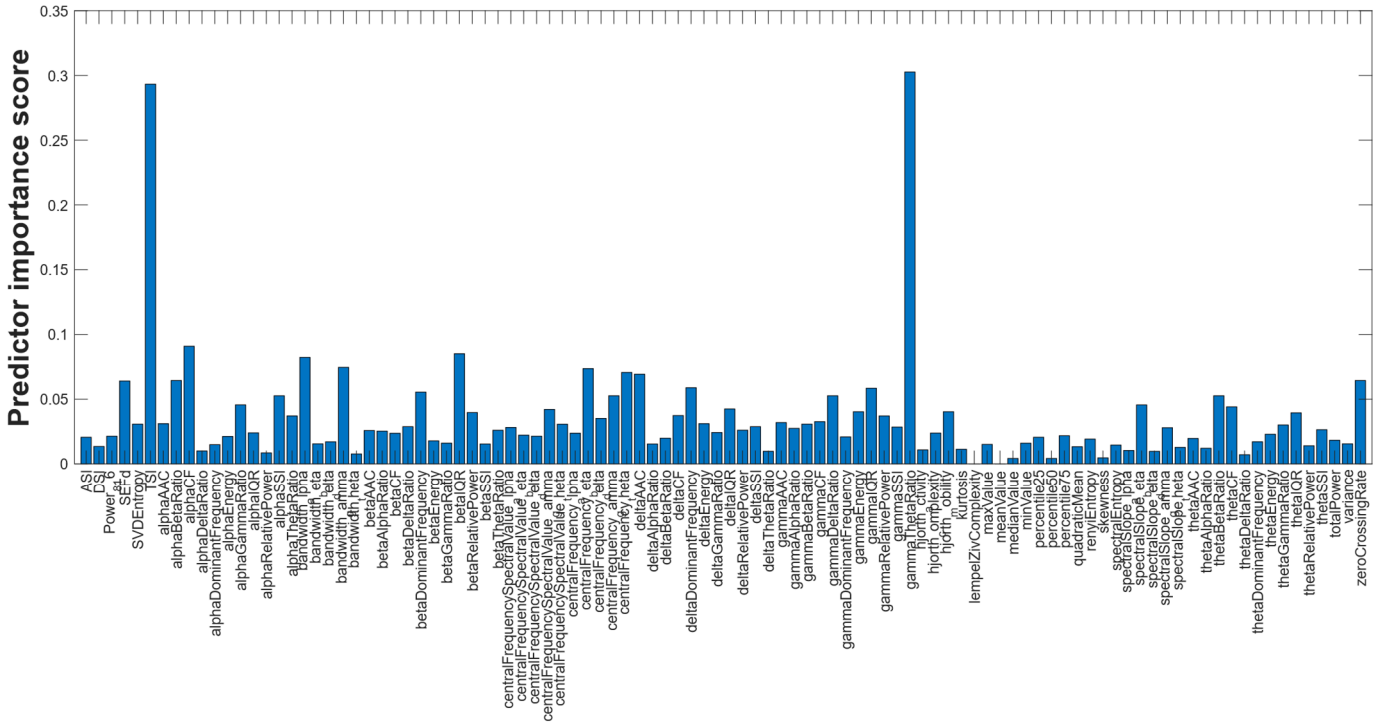


Figure 7. Bar graph representing the predictor importance score for the selected EEG features.

Table 12. Rank of EEG features obtained by the mRMR algorithm with a score higher than 0.01.

Rank	Feature	Score	Rank	Feature	Score
1.	gammaThetaRatio	0.3027	46.	gammaAlphaRatio	0.0275
2.	TSI	0.2933	47.	thetaSSI	0.0265
3.	alphaCF	0.0909	48.	deltaRelativePower	0.0261
4.	betaIQR	0.0851	49.	betaThetaRatio	0.0260
5.	bandwidth_alpha	0.0823	50.	betaAAC	0.0258
6.	bandwidth_gamma	0.0746	51.	betaAlphaRatio	0.0253
7.	centralFrequency_beta	0.0736	52.	deltaGammaRatio	0.0243
8.	centralFrequency_theta	0.0707	53.	alphaIQR	0.0240
9.	deltaAAC	0.0693	54.	hjorth_complexity	0.0238
10.	alphaBetaRatio	0.0645	55.	centralFrequency_alpha	0.0237

Table 12. *Cont.*

Rank	Feature	Score	Rank	Feature	Score
11.	zeroCrossingRate	0.0645	56.	betaCF	0.0237
12.	SEFd	0.0641	57.	thetaEnergy	0.0230
13.	deltaDominantFrequency	0.0589	58.	centralFrequencySpectralValue_beta	0.0222
14.	gammaIQR	0.0585	59.	percentile75	0.0218
15.	betaDominantFrequency	0.0555	60.	Power_8_16	0.0214
16.	gammaDeltaRatio	0.0527	61.	centralFrequencySpectralValue_delta	0.0214
17.	alphaSSI	0.0527	62.	alphaEnergy	0.0212
18.	thetaBetaRatio	0.0527	63.	gammaDominantFrequency	0.0209
19.	centralFrequency_gamma	0.0527	64.	percentile25	0.0206
20.	alphaGammaRatio	0.0456	65.	ASI	0.0206
21.	spectralSlope_beta	0.0456	66.	deltaBetaRatio	0.0199
22.	thetaCF	0.0441	67.	thetaAAC	0.0197
23.	deltaIQR	0.0425	68.	renyiEntropy	0.0192
24.	centralFrequencySpectralValue_gamma	0.0421	69.	totalPower	0.0183
25.	hjorth_mobility	0.0403	70.	betaEnergy	0.0178
26.	gammaEnergy	0.0403	71.	thetaDominantFrequency	0.0171
27.	betaRelativePower	0.0397	72.	bandwidth_delta	0.0171
28.	thetaIQR	0.0395	73.	betaGammaRatio	0.0161
29.	deltaCF	0.0374	74.	minValue	0.0160
30.	alphaThetaRatio	0.0370	75.	variance	0.0155
31.	gammaRelativePower	0.0370	76.	bandwidth_beta	0.0155
32.	centralFrequency_delta	0.0351	77.	deltaAlphaRatio	0.0154
33.	gammaCF	0.0326	78.	betaSSI	0.0154
34.	gammaAAC	0.0320	79.	maxValue	0.0151
35.	deltaEnergy	0.0311	80.	alphaDominantFrequency	0.0149
36.	alphaAAC	0.0311	81.	spectralEntropy	0.0146
37.	gammaBetaRatio	0.0307	82.	thetaRelativePower	0.0140
38.	SVDEntropy	0.0307	83.	DSI	0.0135
39.	centralFrequencySpectralValue_theta	0.0307	84.	quadraticMean	0.0133
40.	thetaGammaRatio	0.0301	85.	spectralSlope_theta	0.0128
41.	betaDeltaRatio	0.0288	86.	thetaAlphaRatio	0.0121
42.	deltaSSI	0.0288	87.	kurtosis	0.0113
43.	gammaSSI	0.0285	88.	hjorth_activity	0.0108
44.	centralFrequencySpectralValue_alpha	0.0281	89.	spectralSlope_alpha	0.0104
45.	spectralSlope_gamma	0.0280	90.	alphaDeltaRatio	0.0100

Figure 8 illustrates the percentage of total variance explained by the PCs (x-axis). By setting a 98% variance threshold, the feature set dimensionality was reduced from 90 features to 56 PCs, thereby achieving the best performance of the developed sleep staging algorithm, as detailed in Section 3.4. The selected PCs represent the optimal combination of the original features, ensuring maximum retention of useful information for sleep staging. The resulting 56 PCs were then integrated into the multimodal feature set used to train the proposed DL model for five-class sleep staging.

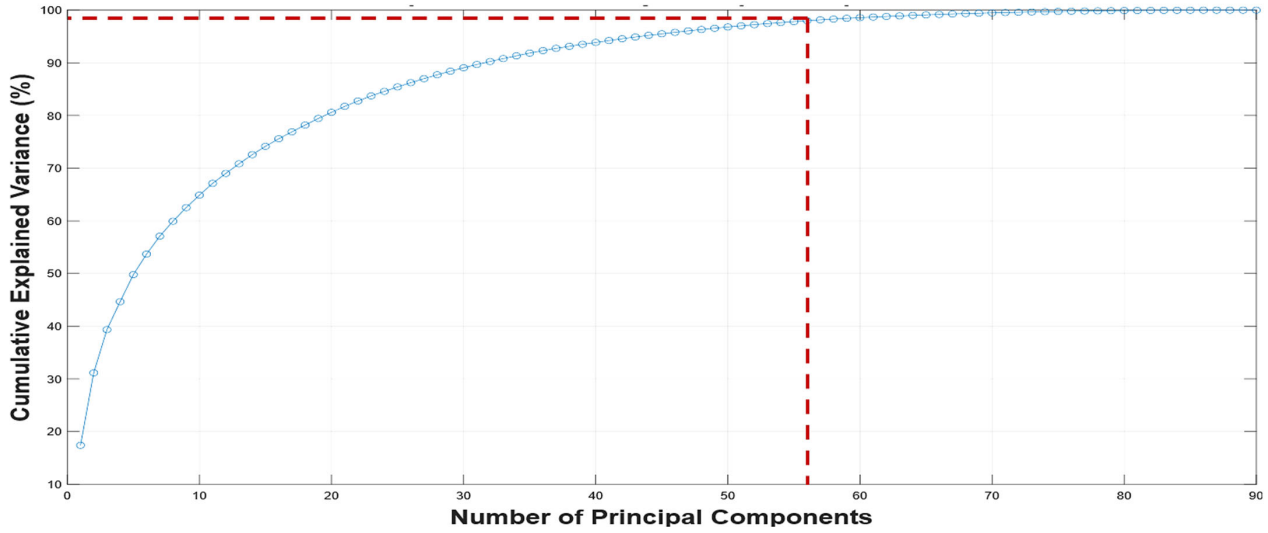


Figure 8. Cumulative explained variance as a function of the number of principal components. The 56th principal component, which accounts for 98% of the total variance, is highlighted in red.

The same selection procedure used for EEG features was also applied to the EOG features of the training set. The mRMR algorithm ranked each feature in the full EOG set by relevance and redundancy, identifying the most informative and least redundant features. Applying a threshold of 0.01 reduced the EOG feature set from 14 to 9 features (Table 13). Figure 9 shows the resulting mRMR scores for the selected EOG features.

Table 13. Rank of EOG features obtained by the mRMR algorithm with a score higher than 0.01.

Rank	Feature	Score	Rank	Feature	Score
1.	PowerBetaGamma	0.03645	6.	MeanBlinkDuration	0.01877
2.	BlinkFrequency	0.03049	7.	StdBlinkDuration	0.01295
3.	PowerAt0_5Hz	0.02496	8.	CurtBlinkDuration	0.01295
4.	MeanBlinkAmplitude	0.02109	9.	SEM_Power	0.01048
5.	PowerDeltaTheta	0.01895			

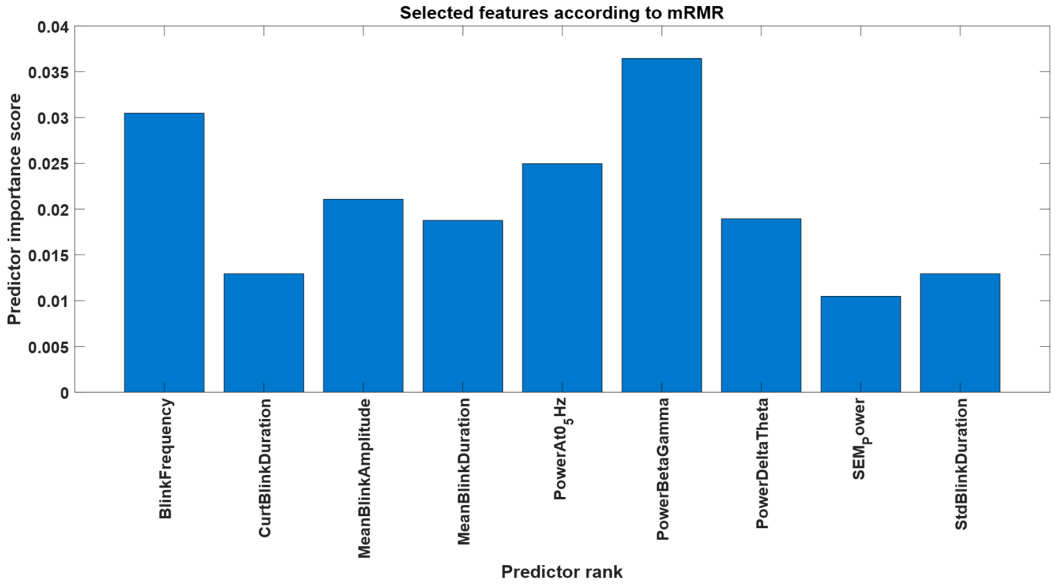


Figure 9. Bar graph representing the predictor importance score for the selected EOG features.

Afterward, PCA was applied to extract the PCs. PCA yielded an optimized reduced feature set comprising 8 principal components, accounting for 98% of the cumulative variance and preserving most of the informative content of the original EOG features. Figure 10 illustrates the cumulative explained variance as a function of the number of PCs for EOG. The optimized PC set was then used to train the DL algorithm for sleep staging, complementing the EEG-based features in the multimodal analysis.

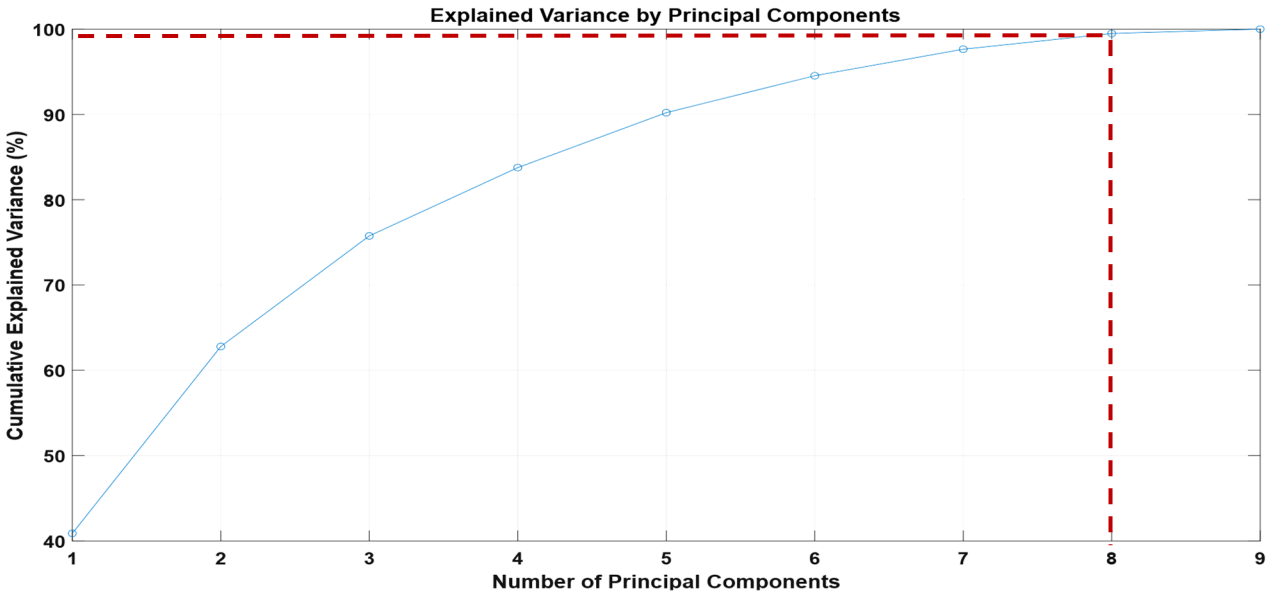


Figure 10. Cumulative explained variance as a function of the number of principal components. The 8th principal component, which accounts for 98% of the total variance, is highlighted in red.

4.2. Training and Testing of a Five-Class Sleep Staging Algorithm

This section presents the performance of the 5-class sleep staging algorithm, trained and tested with different feature sets containing features/PCs from different biosignals, from only EEG PCs (to evaluate the contribution of EEG alone) up to the complete multimodal set (EEG PCs, EOG PCs, and PPG-derived respiratory and cardiac features). Finally, the results without EEG features (using only EOG PCs and PPG features) show lower model performance, demonstrating the high contribution of EEG features.

As discussed in Section 3.4, the proposed algorithm combines two sequential three-class LSTM-based models with an attention mechanism for five-class sleep staging. The first model classifies sequences into Wake, REM, and NREM; the second further distinguishes N1, N2, and N3 within NREM sequences. Despite class imbalance (particularly limited N1 and N3), augmentation strategies (Section 3.5) enabled reliable performance across all classes. This hierarchical framework leverages model specialization for robust five-class staging. First, the WRN and NREM classifiers were trained and tested using the feature sets containing only the 56 EEG PCs. Figure 11 reports the absolute and normalized confusion matrices of the WRN classifier evaluated on the test set; Table 14 summarizes the corresponding evaluation metrics (precision, recall, F1-score, and total accuracy). In particular, the WRN classifier achieves a total accuracy of 89.9%.

Also, the absolute and normalized confusion matrices for the three-class NREM classifier, evaluated on the test set, are shown in Figure 12. In addition, Table 15 summarizes the corresponding evaluation metrics (precision, recall, F1-score, and total accuracy). In particular, the NREM classifier achieves a total accuracy of 91.6%.

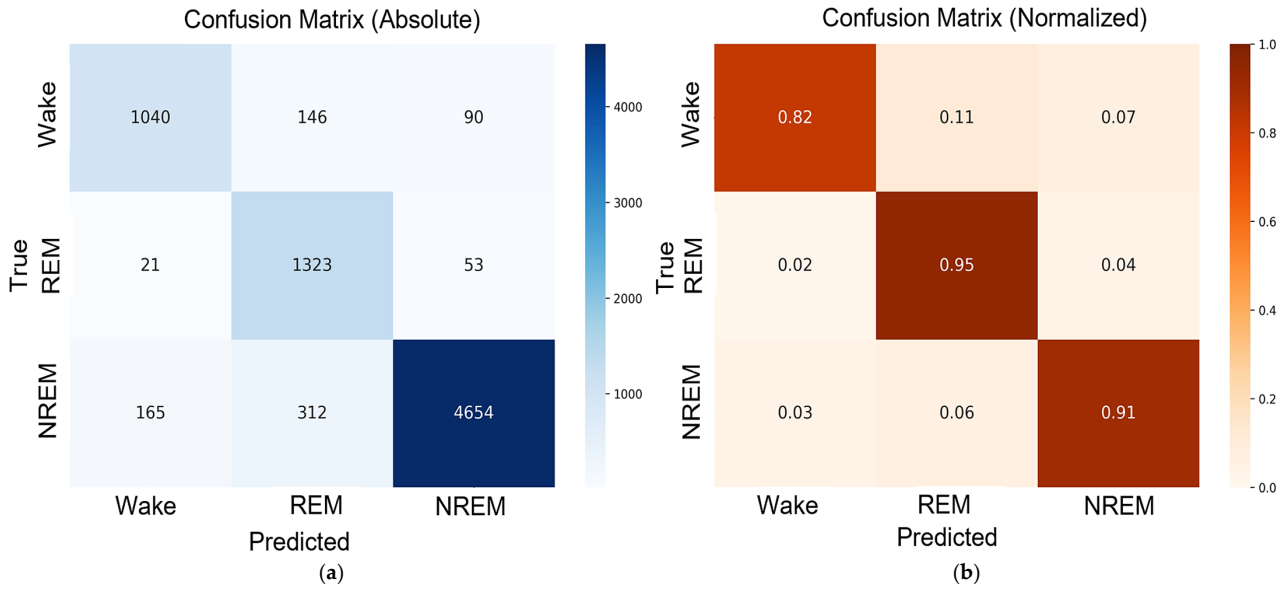


Figure 11. Absolute (**a**) and normalized (**b**) confusion matrices for the WRN classifier trained and tested using the feature set constituted by the 56-PCs EEG (support refers to 50 s-sequences).

Table 14. Evaluation metrics of the 3-class WRN classifier trained and tested on the feature set of 56-PCs EEG (support refers to 50 s-sequences).

	Precision	Recall	F1-Score	Support for Each Class [Sequences]
Wake	0.848 ± 0.013	0.815 ± 0.013	0.831± 0.012	1276
REM	0.743 ± 0.013	0.947 ± 0.011	0.833 ± 0.011	1397
NREM	0.970 ± 0.005	0.907 ± 0.004	0.938 ± 0.007	5131
Macro avg	0.854 ± 0.007	0.890 ± 0.007	0.867 ± 0.006	7804
Weighted avg	0.910 ± 0.004	0.899 ± 0.005	0.901 ± 0.002	7804
Total accuracy	0.899 (i.e., 89.9 ± 0.35%)	Cohen's κ	0.811	7804

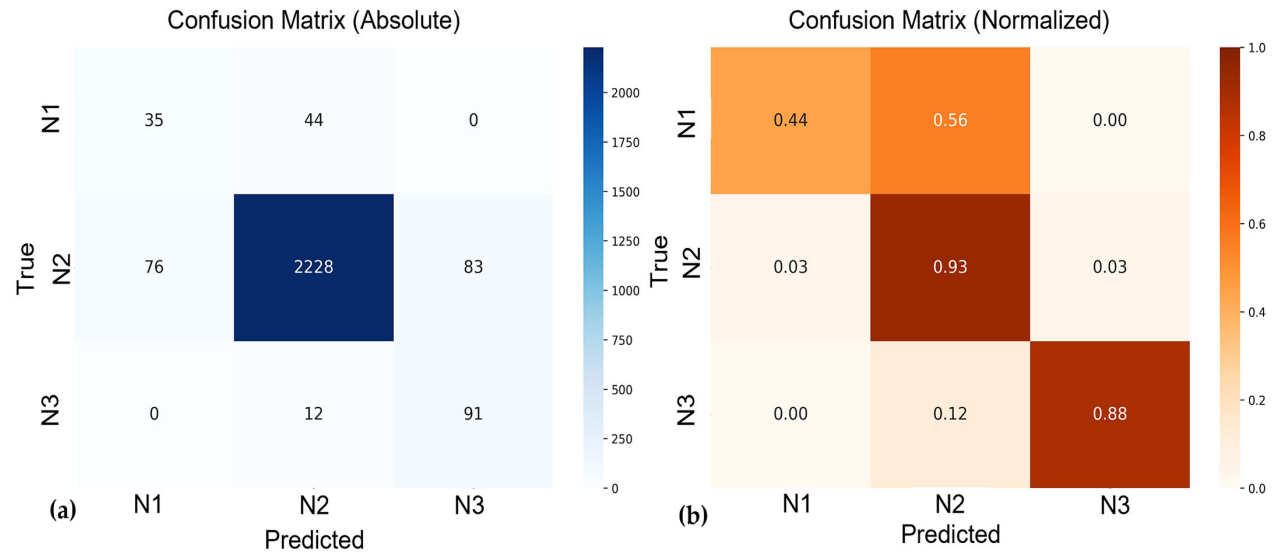


Figure 12. Absolute (**a**) and normalized (**b**) confusion matrices for the NREM classifier trained and tested using the feature set of 56-PCs EEG (support refers to 50 s-sequences).

Table 15. Evaluation metrics of the 3-class NREM classifier trained and tested on the feature set of 56-PCs EEG (support refers to 50 s-sequences).

	Precision	Recall	F1-Score	Support for Each Class [Sequences]
N1	0.315 ± 0.043	0.443 ± 0.062	0.368 ± 0.065	79
N2	0.975 ± 0.009	0.933 ± 0.008	0.954 ± 0.003	2387
N3	0.523 ± 0.054	0.883 ± 0.021	0.657 ± 0.036	103
Macro avg	0.605 ± 0.014	0.753 ± 0.021	0.660 ± 0.023	2569
Weighted avg	0.937 ± 0.006	0.916 ± 0.007	0.924 ± 0.011	2569
Total accuracy	0.916 (i.e., 91.6 ± 0.71%)	Cohen's κ	0.507	2569

The predictions from both three-class classifiers were combined in a cascade framework: the first step classified Wake, REM, and NREM stages using the first model, and upon detecting an NREM epoch, the second model classified it into N1, N2, or N3 stages. The confusion matrices (absolute and normalized) of the two-stage five-class classifier on the test set are reported in Figure 13, along with the corresponding evaluation metrics in Table 16. To quantify the variability of the reported performance metrics, a subject-level bootstrap resampling procedure was performed using 200 bootstrap iterations, applied to the 7-subject test set. At each iteration, the independent test set was resampled by randomly selecting complete subjects with replacement, thereby preserving the hierarchical structure of the data and the temporal dependencies among the sequences belonging to the same subject. The performance metrics were computed for each bootstrap realization, and the values reported in Tables 14–22 are expressed as the mean ± standard deviation across the 200 bootstrap iterations. By resampling complete subjects, the bootstrap estimates provide uncertainty measures that better reflect between-subject variability.

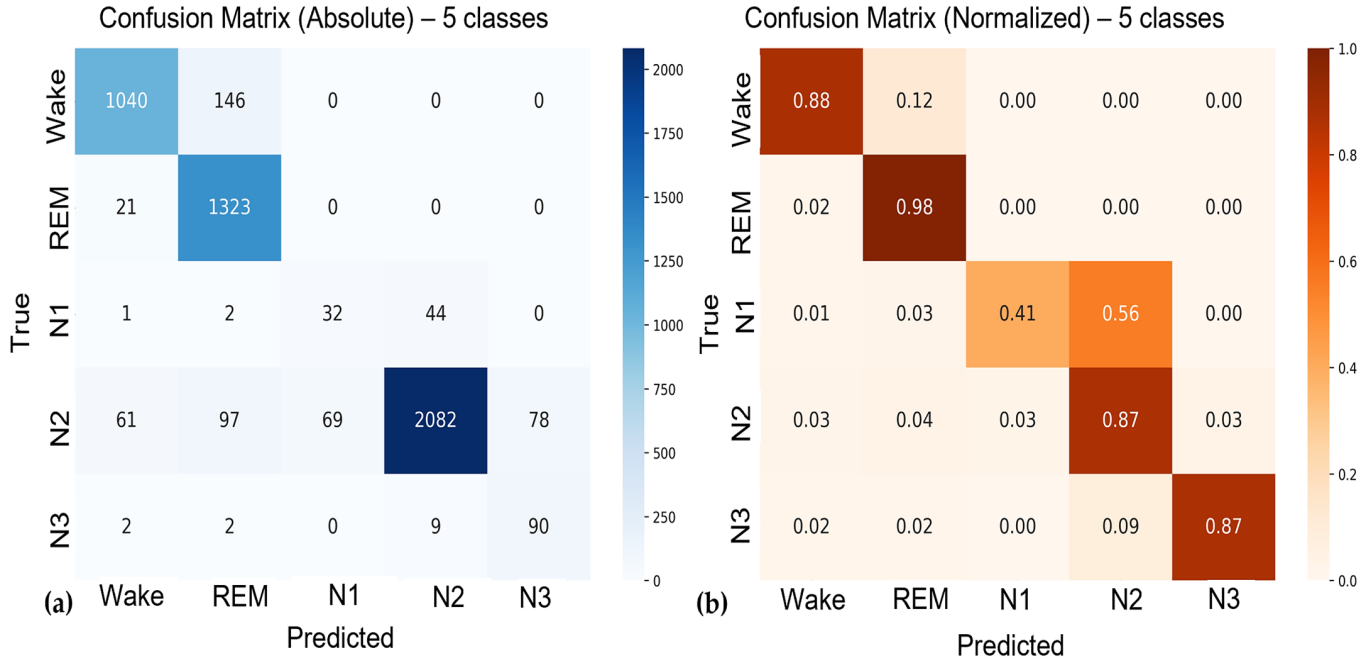


Figure 13. Absolute (**a**) and normalized (**b**) confusion matrices for the 5-class classifier (feature set of 56-PCs EEG, support refers to 50 s-sequences).

Table 16. Evaluation metrics of the 5-class classifier trained and tested on the feature set of 56-PCs EEG (support refers to 50 s-sequences).

	Precision	Recall	F1-Score	Support for Each Class [Sequences]
Wake	0.924 ± 0.012	0.815± 0.012	0.866 ± 0.012	1276
REM	0.843 ± 0.012	0.947 ± 0.008	0.892 ± 0.015	1397
N1	0.317 ± 0.067	0.405 ± 0.076	0.356 ± 0.061	79
N2	0.975 ± 0.006	0.872 ± 0.012	0.921 ± 0.006	2387
N3	0.536 ± 0.022	0.874 ± 0.024	0.664 ± 0.022	103
Macro avg	0.719 ± 0.014	0.783 ± 0.021	0.740 ± 0.017	5242
Weighted avg	0.909 ± 0.017	0.871 ± 0.005	0.886 ± 0.005	5242
Total accuracy	0.871 (i.e., 87.1 ± 0.43%)	Cohen's κ	0.844	5242

Table 17. Evaluation metrics of the WRN classifier trained and tested on the complete feature set (support refers to 50 s-sequences).

	Precision	Recall	F1-Score	Support for Each Class [Sequences]
Wake	0.880 ± 0.014	0.863 ± 0.011	0.871 ± 0.013	1276
REM	0.775 ± 0.013	0.947 ± 0.007	0.852 ± 0.009	1397
NREM	0.973 ± 0.005	0.919 ± 0.006	0.945 ± 0.004	5131
Macro avg	0.876 ± 0.006	0.910 ± 0.008	0.890 ± 0.007	7804
Weighted avg	0.922 ± 0.005	0.915 ± 0.007	0.917 ± 0.005	7804
Total accuracy	0.915 (i.e., 91.5 ± 0.23%)	Cohen's κ	0.838	7804

Table 18. Evaluation metrics of the NREM classifier trained and tested with the complete feature set (support refers to 50 s-sequences).

	Precision	Recall	F1-Score	Support for Each Class [Sequences]
N1	0.294 ± 0.033	0.380 ± 0.041	0.331 ± 0.036	79
N2	0.974 ± 0.010	0.926 ± 0.009	0.950 ± 0.007	2387
N3	0.472 ± 0.021	0.903 ± 0.023	0.620 ± 0.033	103
Macro avg	0.580 ± 0.023	0.736 ± 0.031	0.634 ± 0.027	2569
Weighted avg	0.933 ± 0.008	0.909 ± 0.009	0.917 ± 0.008	2569
Total accuracy	0.909 (i.e., 90.9 ± 0.51%)	Cohen's κ	0.476	2569

Table 19. Evaluation metrics of the 5-class classifier trained and tested on the complete feature set (support refers to 50 s-sequences).

	Precision	Recall	F1-Score	Support for Each Class [Sequences]
Wake	0.939 ± 0.016	0.863 ± 0.013	0.900 ± 0.0010	1276
REM	0.872 ± 0.008	0.947 ± 0.011	0.908 ± 0.008	1397
N1	0.305 ± 0.074	0.367 ± 0.069	0.333 ± 0.064	79

Table 19. *Cont.*

	Precision	Recall	F1-Score	Support for Each Class [Sequences]
N2	0.972 ± 0.012	0.871 ± 0.0011	0.919 ± 0.009	2387
N3	0.484 ± 0.041	0.893 ± 0.037	0.628 ± 0.031	103
Macro avg	0.715 ± 0.015	0.788 ± 0.019	0.737 ± 0.017	5242
Weighted avg	0.918 ± 0.006	0.882 ± 0.003	0.897 ± 0.005	5242
Total accuracy	0.882 (i.e., 88.2 ± 0.34%)	Cohen's κ	0.857	5242

Table 20. Evaluation metrics of the WRN classifier on the feature set comprising 8 EOG PCs, 2 respiratory features, and 8 cardiac features from the PPG signal (excluding the 56 EEG PCs, support refers to 50 s-sequences).

	Precision	Recall	F1-Score	Support for Each Class [Sequences]
Wake	0.488 ± 0.015	0.607 ± 0.018	0.541 ± 0.020	1276
REM	0.436 ± 0.012	0.546 ± 0.017	0.485 ± 0.013	1397
NREM	0.834 ± 0.008	0.727 ± 0.007	0.777 ± 0.008	5131
Macro avg	0.586 ± 0.009	0.626 ± 0.006	0.601 ± 0.009	7804
Weighted avg	0.706 ± 0.004	0.675 ± 0.008	0.686 ± 0.007	7804
Total accuracy	0.675 (i.e., 67.5 ± 0.41%)	Cohen's κ	0.409	7804

Table 21. Evaluation metrics of the 3-class NREM classifier on the feature set comprising 8 EOG PCs, 2 respiratory features, and 8 cardiac features from the PPG signal (excluding the 56 EEG PCs, support refers to 50 s-sequences).

	Precision	Recall	F1-Score	Support for Each Class [Sequences]
N1	0.135 ± 0.044	0.089 ± 0.047	0.107 ± 0.040	79
N2	0.944 ± 0.007	0.820 ± 0.0011	0.878 ± 0.009	2387
N3	0.095 ± 0.031	0.408 ± 0.041	0.154 ± 0.035	103
Macro avg	0.391 ± 0.029	0.439 ± 0.028	0.379 ± 0.025	2569
Weighted avg	0.885 ± 0.0010	0.781 ± 0.014	0.825 ± 0.0011	2569
Total accuracy	0.781 (i.e., 78.1 ± 0.82%)	Cohen's κ	0.097	2569

Table 22. Evaluation metrics of the 5-class classifier on the feature set comprising 8 EOG PCs, 2 respiratory features, and 8 cardiac features from the PPG signal (excluding the 56 EEG PCs, support refers to 50 s-sequences).

	Precision	Recall	F1-Score	Support for Each Class [Sequences]
Wake	0.606 ± 0.021	0.607 ± 0.018	0.606 ± 0.015	1276
REM	0.470 ± 0.017	0.546 ± 0.015	0.505 ± 0.010	1397
N1	0.156 ± 0.077	0.063 ± 0.068	0.090 ± 0.073	79
N2	0.960 ± 0.011	0.508 ± 0.015	0.664 ± 0.012	2387

Table 22. *Cont.*

	Precision	Recall	F1-Score	Support for Each Class [Sequences]
N3	0.030 ± 0.043	0.087 ± 0.041	0.044 ± 0.037	103
Macro avg	0.444 ± 0.019	0.362 ± 0.022	0.382 ± 0.020	5242
Weighted avg	0.713 ± 0.009	0.527 ± 0.010	0.587 ± 0.009	5242
Total accuracy	0.527 (i.e., 52.7 ± 0.59%)	Cohen's κ	0.457	5242

The first task involved training and testing the WRN classifier using a feature set composed exclusively of EEG-derived PCs. The WRN classifier achieved 89.9% total accuracy and a 0.901 macro-average F1-score on the test set (as shown in Table 14). In detail, the Wake stage achieved an F1-score of 0.831, with 0.848 precision and 0.831 recall, indicating reliable wakefulness detection, though with some misclassifications.

Despite the known morphological similarity of EEG between the REM stage and Wake and NREM, the model achieved an F1-score of 0.833, a precision of 0.743, and a recall of 0.947. The NREM stage also showed stable performance, with an F1-score of 0.938, supported by high precision (0.970) and recall (0.938), likely influenced by its strong representation in the dataset. Using the same feature set, the three-class NREM classifier achieved 91.6% total accuracy and a macro-average F1-score of 0.660 in discriminating among N1, N2, and N3 stages (as reported in Table 15). To mitigate class imbalance, data augmentation was applied to N1 and N3 classes, which are typically underrepresented in the dataset (Figure 6b). Despite these measures, the identification of the N1 stage remained the most challenging, achieving an F1-score of 0.368 (precision 0.315, recall 0.443), due to its transitional nature and limited representation in the dataset. The N3 stage also showed modest performance, with a precision of 0.523, a recall of 0.883, and an F1-score of 0.660, indicating that deep sleep is easier to identify than N1. For the N2 stage, which typically accounts for the majority of total sleep time, the model performed well, achieving balanced precision (0.975), recall (0.933), and an F1-score value of 0.954. The overall results of the sleep-staging algorithm, which cascades two three-class classifiers (Table 16), demonstrate that reliable multi-class classification can be achieved using only frontal EEG signals. The total accuracy of 87.1% and a macro-average F1-score of 0.740 confirm the effectiveness of the five-class sleep-staging model, due to the reduced feature set obtained via mRMR and PCA.

Regarding the Wake stage, the LSTM network achieved an F1-score of 0.866, with a precision of 0.924 and a recall of 0.815, indicating consistent performance in wake detection. The REM stage achieved an F1-score of 0.892, with recall (0.943) and precision (0.843), showing good REM detection capability. Among NREM sub-stages, N1 remained the most challenging class, with an F1-score of 0.664, resulting in reduced precision (0.317) and recall (0.405). N1 consistently emerged as the most challenging stage across all configurations, attributable to several factors. First, severe class imbalance (N1 accounts for only 2–5% of total sleep time) limits the diversity of training examples despite targeted augmentation, hindering the model's ability to establish robust boundaries against adjacent Wake and N2 stages. Second, N1 is inherently a transitional stage whose physiological characteristics overlap with those of both Wake and N2. Because the BOAS dataset provides annotations only at the standard 30 s resolution, each 5 s sub-epoch inherited the label of its parent epoch, potentially introducing mixed-stage physiological patterns and increasing intra-class variability. Furthermore, the majority-voting procedure used to assign labels to 10-epoch sequences may reduce the influence of isolated N1 occurrences. Third, frontal derivation sensitivity is inherently limited for N1, since vertex sharp waves (a hallmark of this stage)

are maximal over central derivations and are therefore only partially captured by the frontal channel employed here.

Future work should explore multi-resolution architectures, the addition of central derivations, and loss functions that explicitly penalize N1-Wake and N1-N2 confusions, as well as explicit modeling of stage transitions, given the strongly context-dependent nature of N1 occurrence. Also, the N2 stage achieved solid performance (0.921 F1-score), supported by high precision (0.975) and slightly lower recall (0.881). The N3 stage shows moderate reliability, with an F1-score of 0.664, driven by high recall (0.874) despite moderate precision (0.536). These results indicate that, even with a reduced set of EEG features, deep sequential models can achieve good sleep staging performance. However, classifying the transitional stage N1 remains a critical challenge due to its limited representation in the dataset. Finally, for the five-class sleep staging model, the corresponding Cohen's kappa for the EEG-only configuration was 0.844, indicating very good agreement. For the intermediate classifiers, kappa values were 0.811 (WRN) and 0.507 (NREM), confirming robust performance at both stages of the hierarchical pipeline.

To assess the impact of a different feature set, the WRN and NREM classifiers were trained and tested using the complete feature set, including 56 PCs of EEG, eight PCs of EOG, two respiratory features and eight cardiac features from the PPG signal. Figure 14 reports the absolute and normalized confusion matrices of the WRN classifier evaluated on the test set, with the corresponding evaluation metrics (precision, recall, F1-score, and total accuracy). The WRN classifier achieves a total accuracy of 91.5% (Table 17).

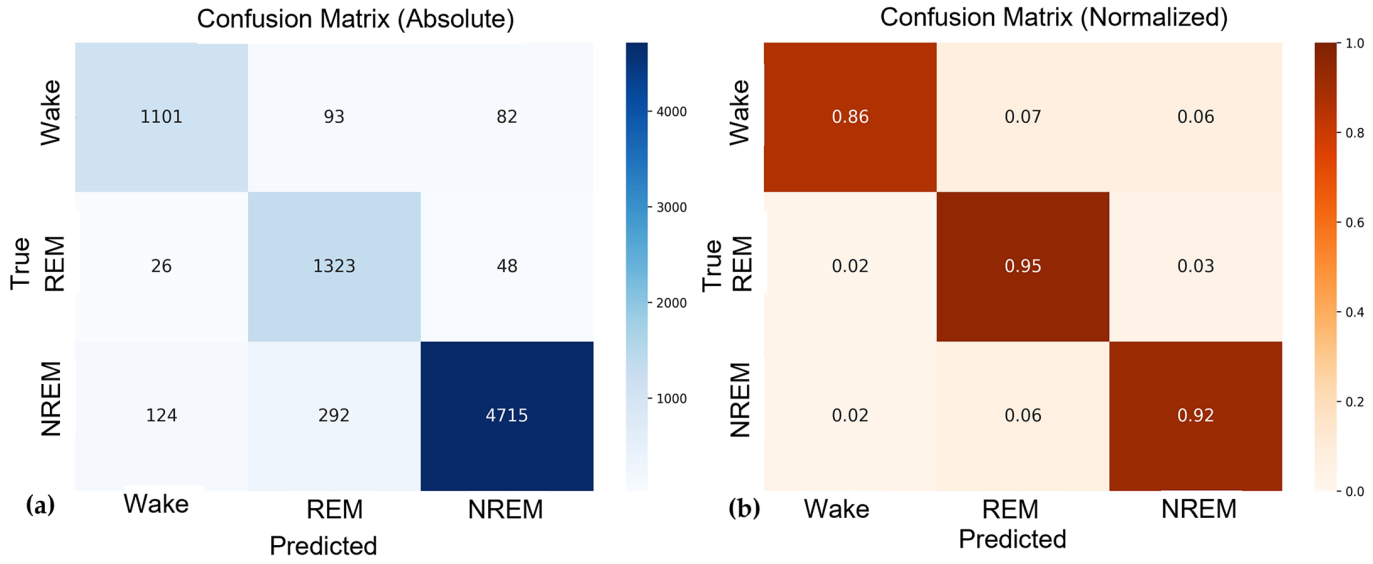


Figure 14. Absolute (**a**) and normalized (**b**) confusion matrices for the WRN classifier with the complete feature set (support refers to 50 s-sequences).

Also, the absolute and normalized confusion matrices for the three-class NREM classifier, evaluated on the test set, are shown in Figure 15. In addition, Table 18 summarizes the corresponding evaluation matrices (precision, recall, F1-score, and total accuracy). In particular, the NREM classifier achieves a total accuracy of 90.9%.

The two three-class classifiers were combined in a cascade to form a five-class classifier. The confusion matrices (absolute and normalized) of the two-stage five-class classifier on the test set are reported in Figure 16, along with the corresponding evaluation metrics in Table 19. Also in Tables 17–19, the reported metrics correspond to mean ± standard deviation across 200 subject-level bootstrap realizations.

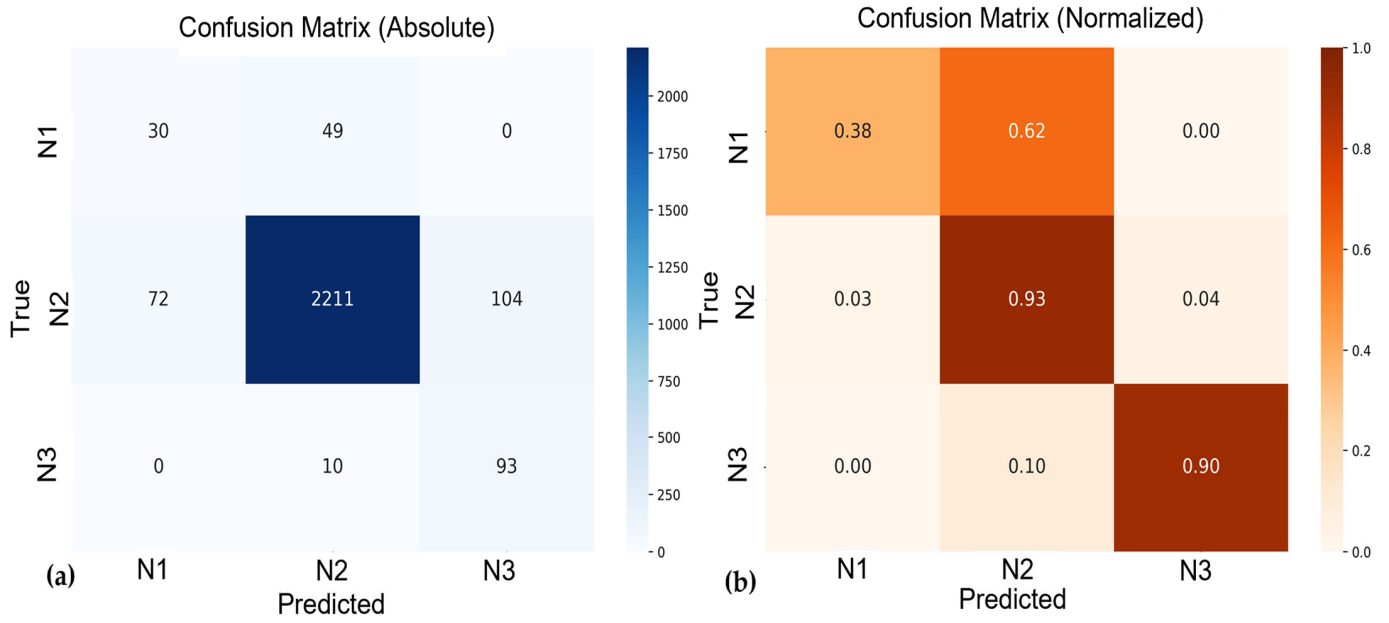


Figure 15. Absolute (**a**) and normalized (**b**) confusion matrices for the NREM classifier with the complete feature set (support refers to 50 s-sequences).

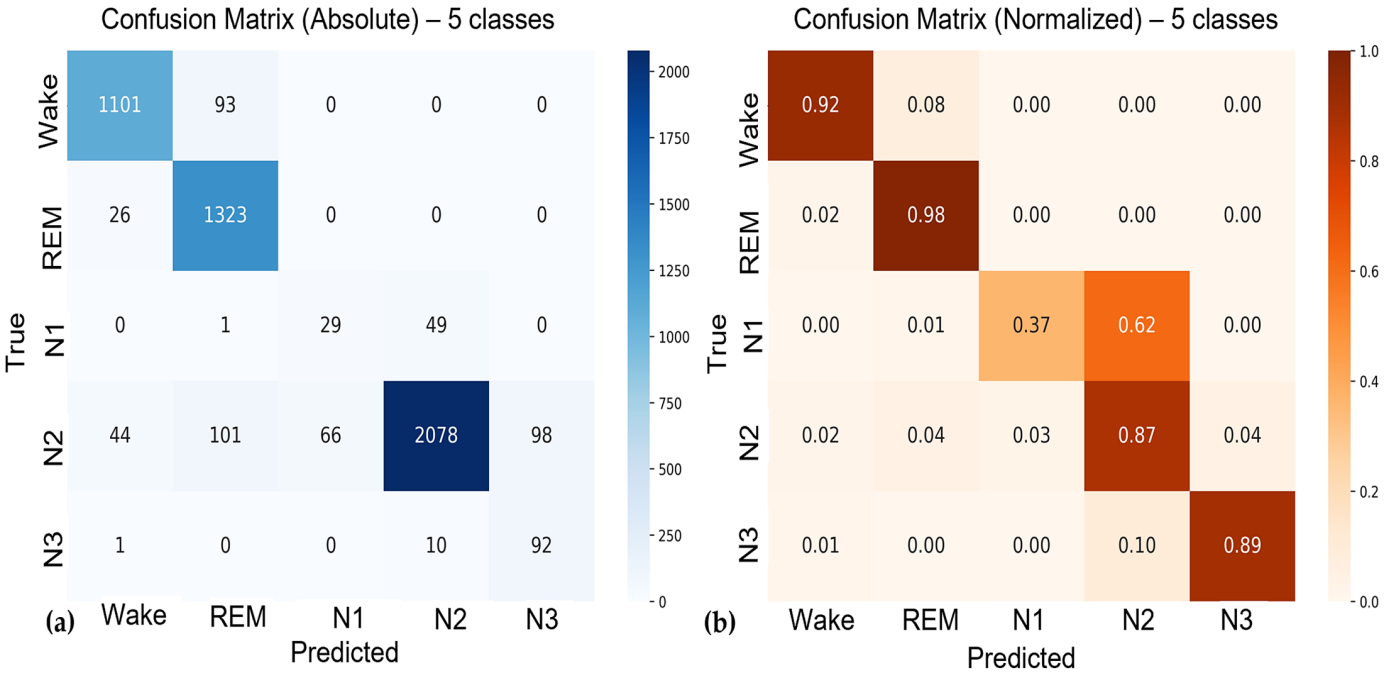


Figure 16. Absolute (**a**) and normalized (**b**) confusion matrices for the 5-class classifier with the complete feature set (support refers to 50 s-sequences).

Beyond the EEG-only setup, the framework was evaluated on a complete multimodal configuration (56 EEG PCs, eight EOG PCs, two respiratory, and eight cardiac PPG features) to assess the incremental contribution of ocular and cardiorespiratory information. Under the complete multimodal feature configuration, the WRN classifier achieved 91.5% total accuracy and a macro-average F1-score of 0.890 (Table 17), indicating a modest contribution of complementary physiological modalities. Wake performance (F1-score 0.871, precision 0.880, recall 0.863) remained comparable to the EEG-only configuration, preserving a good separability of wake epochs. REM stage detection showed a slight improvement (F1-score 0.852, precision 0.775, recall 0.947), likely due to the inclusion of EOG components

that capture Rapid Eye Movements. NREM maintains stable classification performance (F1-score 0.945, precision 0.973, recall 0.919) under the expanded multimodal feature set. Overall, while EEG features dominate WRN discrimination, EOG and cardiorespiratory PPG features provide complementary cues that particularly benefit REM classification.

Using the same multimodal configuration, the three-class NREM classifier achieved 90.9% total accuracy and a macro-average F1-score of 0.634 for N1/N2/N3 classification (as shown in Table 18). The N1 stage achieves performance similar to the EEG-only setup (F1-score 0.331, precision 0.294, recall 0.380), while the N2 stage shows strong, stable classification performance (F1-score 0.950, precision 0.974, recall 0.926). Also, the N3 stage showed moderate deep sleep detection (F1-score 0.620, precision 0.472, recall 0.903) though with a tendency toward overestimation. Overall, the multimodal integration provided modest performance gains, mainly for REM detection.

Finally, the cascaded five-class sleep-staging framework under the multimodal configuration achieved 88.2% accuracy and a macro-average F1-score of 0.737 (Table 19), indicating good overall performance of the complete architecture. The Wake stage achieved an F1-score of 0.900, with a precision of 0.939 and a recall of 0.863, demonstrating reliable wake detection. REM also worked well (F1-score 0.908, recall 0.947, precision 0.872), with few missed epochs and limited false positives. Among the NREM sub-stages, the N1 stage remains the most challenging (F1-score 0.333, precision 0.305, recall 0.367), reflecting its transitional nature and overlap with Wake and N2. The F1-score slightly decreased from 0.356 (EEG-only) to 0.333 (multimodal) after EOG inclusion, suggesting that the EOG information did not improve overall classification performance. However, slow eye movements are generally better captured by EOG than by frontal EEG. The N2 stage, representing the largest portion of sleep time, achieved stable and consistent classification performance (F1-score 0.919, precision 0.972, recall 0.871). N3 stage achieved an F1-score of 0.628, with precision (0.484) and recall (0.893), indicating moderate deep sleep detection with a tendency toward overestimation (higher recall than precision).

Figure 17 presents the adjacent-stage transition-confusion heatmap, a specialized visualization that focuses exclusively on misclassification patterns between physiologically neighboring sleep stages (Wake $\leftrightarrow$ N1, N1 $\leftrightarrow$ N2, N2 $\leftrightarrow$ N3, REM $\leftrightarrow$ Wake, REM $\leftrightarrow$ N2). Unlike standard confusion matrices, this visualization masks non-adjacent transitions to highlight the most clinically relevant confusion boundaries where EEG features exhibit maximum overlap. The heatmap reveals three key findings. First, Wake and REM are well discriminated, with only 8% of Wake epochs misclassified as REM and 2% of REM epochs misclassified as Wake. Second, asymmetric N1-N2 confusion (62% N1 $\rightarrow$ N2, 3% N2 $\rightarrow$ N1) reflects N1's physiological transitional nature, consistent with literature on the inherent difficulty of N1 classification. Third, low confusion for N2 $\leftrightarrow$ N3 (4% N2 $\rightarrow$ N3 and 10% N3 $\rightarrow$ N2) confirms reliable discrimination for remaining adjacent pairs. Overall, the results demonstrate that N1 remains the primary source of classification error, representing a physiologically expected limitation and consistent with state-of-the-art sleep staging performance.

Overall, multimodal integration yielded a modest increase in overall accuracy, while macro-average F1-score remained comparable to the EEG-only configuration. EEG features remain the primary source of discrimination, whereas EOG and cardiorespiratory PPG features may provide complementary information that refines classification decisions. Multimodal fusion may facilitate the discrimination of some stages, particularly REM, although improvements were modest. Also, under the complete multimodal configuration, Cohen's kappa reached 0.857 for the five-class task, while the WRN and NREM classifiers achieved kappa values of 0.838 and 0.476, respectively. Compared with the EEG-only configuration (κ = 0.844), multimodal integration yielded a modest improvement in chance-corrected agreement, attributable to enhanced discrimination of REM stage. The multimodal frame-

work remains lightweight: WRN (1.91 MB), NREM (1.23 MB), cascaded total (3.14 MB). Compared to the EEG-only configuration, the 0.19 MB increase yields a 1.1% improvement in accuracy, achieving the highest accuracy in this study, as reported in Section 5.

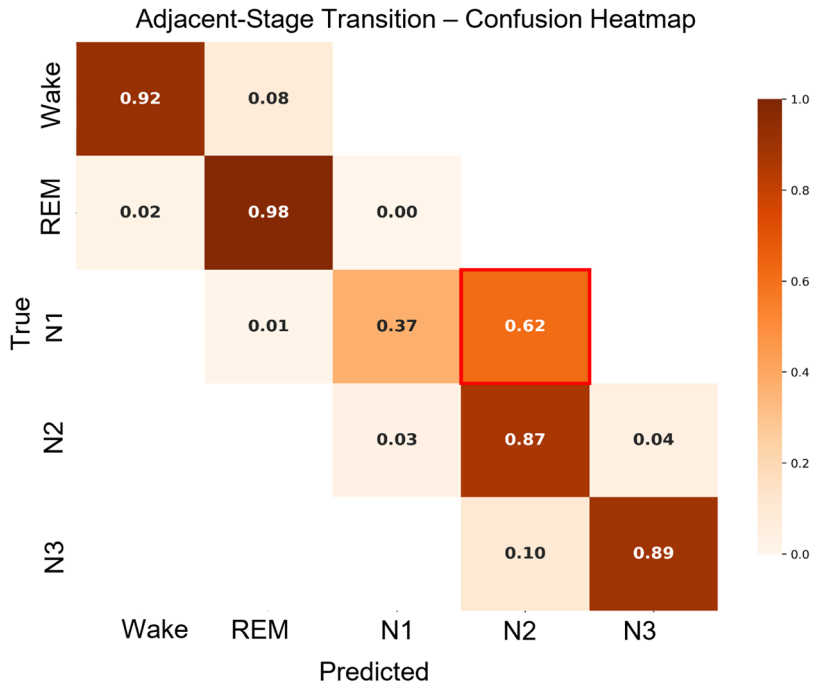


Figure 17. Adjacent-stage transition-confusion heatmap. Values represent misclassification proportions between adjacent sleep stages for the 5-class classifier with complete feature set.

Finally, the WRN and NREM classifiers were trained and tested using a feature set comprising the eight EOG PCs, two respiratory features, and eight cardiac features from the PPG signal (excluding the 56 EEG PCs). Figure 18 shows the absolute and normalized confusion matrices of the WRN classifier evaluated on the test set; Table 20 summarizes the corresponding evaluation metrics (precision, recall, F1-score, and total accuracy). In particular, the WRN classifier achieves a total accuracy of 67.5%.

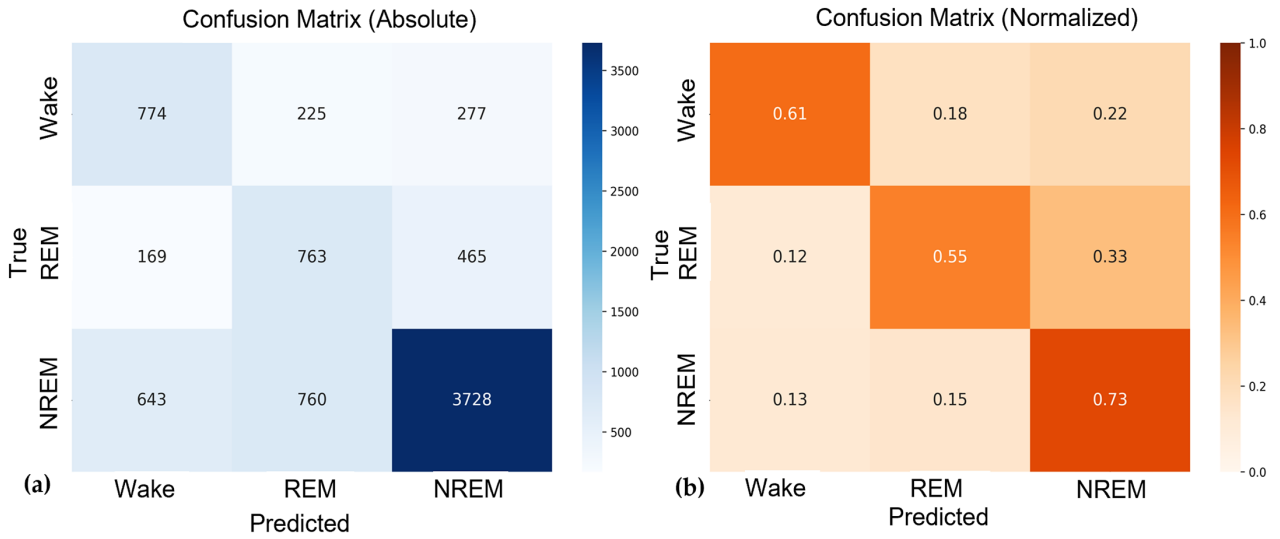


Figure 18. Absolute (**a**) and normalized (**b**) confusion matrices for the WRN classifier using the feature set comprising 8 EOG PCs, 2 respiratory features, and 8 cardiac features from the PPG signal (excluding the 56 EEG PCs, support refers to 50 s-sequences).

Also, the absolute and normalized confusion matrices for the three-class NREM classifier, evaluated on the test set, are shown in Figure 19. In addition, Table 21 summarizes the corresponding evaluation metrics (precision, recall, F1-score, and total accuracy). In particular, the NREM classifier obtains a total accuracy of 78.1%.

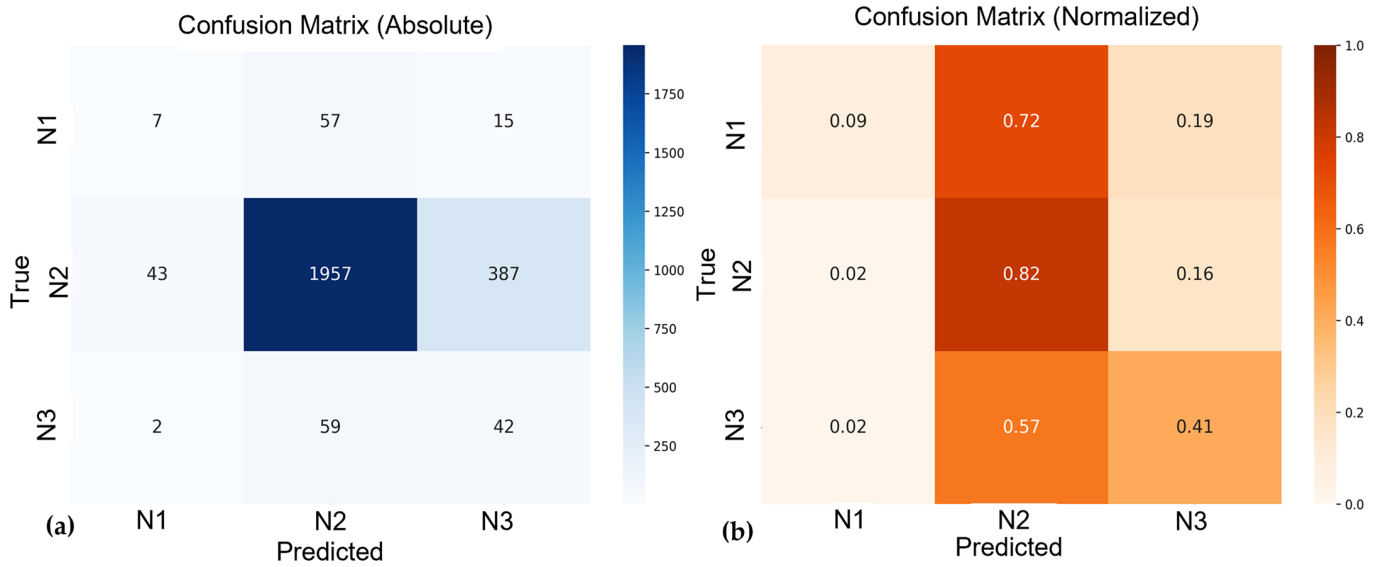


Figure 19. Absolute (**a**) and normalized (**b**) confusion matrices for the NREM classifier using the feature set comprising 8 EOG PCs, 2 respiratory features, and 8 cardiac features from the PPG signal (excluding the 56 EEG PCs, support refers to 50 s-sequences).

The two three-class classifiers were combined in a cascade to form a five-class classifier. The confusion matrices (absolute and normalized) of the two-stage five-class classifier on the test set are reported in Figure 20, along with the corresponding evaluation metrics in Table 22. As in the previous feature set, subject-level bootstrap resampling (200 iterations) was also applied to evaluate the variability of the performance metrics reported in Tables 20–22.

Under the EEG-excluded multimodal configuration (8 EOG PCs, 2 respiratory, and 8 cardiac PPG features), the WRN classifier achieved 67.5% total accuracy and a macro-average F1-score of 0.601 (Table 20). Ocular and cardiorespiratory signals alone contribute to coarse sleep-stage discrimination, though performance remains substantially below EEG-only or full multimodal setups. Wake detection achieved modest performance (F1-score 0.541, precision 0.488, recall 0.607), with many correct identifications but a substantial number of false positives, highlighting the central role of neural signals in accurate wake detection. REM achieved an F1-score of 0.485 (precision 0.436 and recall 0.546). While the recall of 0.546 shows that over half of REM epochs are detected, the relatively low precision (0.436) indicates a substantial number of false positives, suggesting that REM-like ocular and cardiorespiratory patterns are insufficiently distinctive without EEG input.

NREM stage reached an F1-score of 0.777 (precision 0.834, recall 0.727), showing that coarse NREM epochs can be moderately identified using only cardiorespiratory and ocular signals. Overall, reduced performance without EEG confirms that EEG features provide the primary discriminative power for sleep-stage classification. EOG and PPG features provide complementary but insufficient information on their own. The moderate REM recall indicates some discriminative value, but limited specificity leads to more false positives when neural information is excluded.

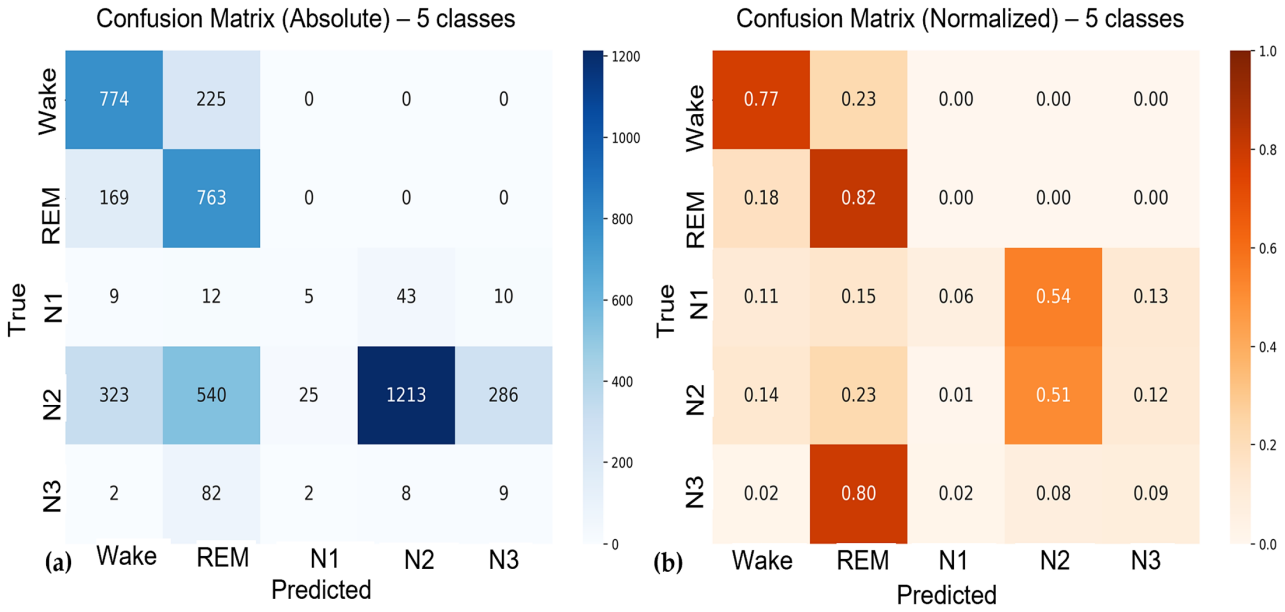


Figure 20. Absolute (**a**) and normalized (**b**) confusion matrices for the 5-class classifier with the feature set comprising 8 EOG PCs, 2 respiratory, and 8 cardiac features from the PPG signal (excluding the 56 EEG PCs, support refers to 50 s-sequences).

Using the same EEG-excluded feature set, the NREM classifier achieved 78.1% total accuracy and a macro-average F1-score of 0.379 (Table 21). N1 performance was poor (F1-score 0.107, precision 0.135, recall 0.089), reflecting the difficulty of identifying transitional sleep without EEG markers. The N2 stage showed good performance, particularly high sensitivity (F1-score 0.878, precision 0.944, recall 0.820). N3 achieved a recall of 0.408 but low precision (0.095), resulting in an F1 of 0.154; this indicates that deep sleep detection is limited in sensitivity but prone to false positives without EEG input. These results indicate that NREM sub-stages carry some discriminative cardiorespiratory and ocular signatures detectable from EOG and PPG alone. However, the lower overall performance compared to the EEG-inclusive feature set highlights the main limitation of the non-EEG setup: while N2 and N3 can still be classified with good performance, N3 and transitional N1 epochs are not reliably identified.

Finally, the cascaded five-class framework using only EOG and PPG-derived features achieved 52.7% accuracy and a macro-average F1-score of 0.382 (Table 22), indicating that ocular and cardiorespiratory signals provide meaningful information for basic sleep-stage discrimination, but optimal performance requires EEG features. Wake detection performed modestly without EEG input (F1-score 0.606, precision 0.606, recall 0.607), though some misclassifications occur. REM achieved an F1-score of 0.505 (precision 0.470, recall 0.546), showing modest performance with both missed epochs and false positives, highlighting the limitations of REM discrimination without neural signals.

Among the NREM sub-stages, N1 performed poorly (F1-score 0.090, precision 0.156, recall 0.063), whereas N2 achieved good precision (0.960) but lower recall (0.508), resulting in an F1-score of 0.664, indicating that N2 is identified with high confidence when predicted (high precision), but some true N2 epochs are missed (lower recall). N3 reached an F1-score of 0.044 (precision 0.030, recall 0.087), showing that deep sleep is rarely detected correctly and is associated with a high number of false positives. Excluding EEG features reduced Cohen's kappa to 0.457 for the five-class task and to 0.409 and 0.097 for the WRN and NREM classifiers, respectively. This kappa value falls within the moderate agreement range, confirming that while EOG and PPG signals carry discriminative information, their standalone classification capacity is significantly constrained compared to EEG-inclusive

configurations. These results show that although EOG and PPG signals contain useful information for sleep-stage classification, the absence of EEG-derived features significantly limits the model's ability to distinguish among the five classes reliably.

To further characterize the proposed cascade sleep classifier, Table 23 reports the subject-wise performance for the EEG-only and multimodal configurations achieved for the seven subjects included in the test set. For each subject, the table reports total accuracy, weighted precision, weighted recall, weighted F1-score, and Cohen's κ, along with the paired difference (multimodal minus EEG-only). Also, the table summarizes the aggregate results, reporting the mean difference and its 95% confidence intervals across subjects, enabling quantification of the inter-subject variability and the overall improvement introduced by the multimodal configuration.

Table 23. Subject-wise performance of EEG-only and multimodal (EEG + EOG + PPG) configurations on the seven held-out test subjects. For each subject, the table reports accuracy, precision, recall, F1-score, and Cohen's κ, along with the paired difference (multimodal minus EEG-only). Bottom rows show the mean difference across subjects and its 95% confidence interval.

Subject	Feature Set	Total Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score	Cohen's κ
72	Only EEG	0.890	0.924	0.891	0.907	0.858
	EEG + EOG + PPG	0.905	0.934	0.913	0.926	0.879
	Difference	+0.015	+0.010	+0.022	+0.019	+0.021
62	Only EEG	0.885	0.920	0.889	0.906	0.848
	EEG + EOG + PPG	0.895	0.922	0.889	0.901	0.842
	Difference	+0.010	+0.002	+0.000	−0.005	−0.006
82	Only EEG	0.855	0.942	0.853	0.892	0.818
	EEG + EOG + PPG	0.870	0.951	0.858	0.898	0.825
	Difference	+0.015	+0.009	+0.005	+0.006	+0.007
88	Only EEG	0.845	0.897	0.842	0.858	0.829
	EEG + EOG + PPG	0.875	0.915	0.878	0.892	0.885
	Difference	+0.030	+0.018	+0.036	+0.034	+0.056
71	Only EEG	0.860	0.875	0.852	0.855	0.830
	EEG + EOG + PPG	0.880	0.892	0.863	0.877	0.850
	Difference	+0.020	+0.017	+0.011	+0.022	+0.020
87	Only EEG	0.865	0.885	0.845	0.859	0.830
	EEG + EOG + PPG	0.885	0.891	0.860	0.868	0.846
	Difference	+0.020	+0.006	+0.015	+0.009	+0.016
49	Only EEG	0.897	0.920	0.925	0.925	0.891
	EEG + EOG + PPG	0.880	0.921	0.913	0.908	0.872
	Difference	−0.017	+0.001	−0.012	−0.017	−0.019
Mean difference		+0.013	+0.009	+0.011	+0.010	+0.014
Confidence Interval		[−0.004, +0.030]	[−0.002, +0.020]	[−0.004, +0.026]	[−0.006, +0.026]	[−0.005, +0.033]

The inclusion of EOG and PPG signals resulted in a modest performance improvement compared with the EEG-only configuration. Across the seven test subjects, the multimodal configuration showed a mean improvement of 1.3% in total accuracy, with

a confidence interval crossing zero (95% CI: −0.4% to +3%), indicating that the improvement is not statistically significant (Table 23). Six of seven subjects showed improvement in accuracy, while one subject (Subject 49) exhibited a degradation of 1.7%, highlighting inter-subject variability. Overall, these results indicate that multimodal information may provide complementary features to EEG, although the observed improvement is modest and not statistically significant. The multimodal benefit should therefore be characterized as complementary rather than a general increase in robustness.

4.3. Contribution of EOG and PPG Signals

To further investigate the individual contribution of the non-EEG signal modalities, an additional feature-group ablation study was conducted. The proposed classifier was independently trained and evaluated using only the EOG PCs and only the PPG-derived features, while maintaining the same training and testing protocol adopted throughout this work. The obtained results are summarized in Tables 24 and 25.

Table 24. Evaluation metrics of the 5-class classifier on the feature set comprising 8 EOG PCs (support refers to 50 s-sequences).

	Precision	Recall	F1-Score	Support for Each Class [Sequences]
Wake	0.523 ± 0.017	0.589 ± 0.020	0.554 ± 0.015	1276
REM	0.741 ± 0.020	0.440 ± 0.021	0.552 ± 0.013	1397
N1	0.220 ± 0.042	0.114 ± 0.055	0.150 ± 0.035	79
N2	0.957 ± 0.013	0.677 ± 0.019	0.793 ± 0.019	2387
N3	0.184 ± 0.028	0.612 ± 0.045	0.283 ± 0.029	103
Macro avg	0.525 ± 0.021	0.486 ± 0.023	0.466 ± 0.018	5242
Weighted avg	0.767 ± 0.015	0.583 ± 0.015	0.651 ± 0.014	5242
Total accuracy	0.583 (i.e., 58.3 ± 0.77%)	Cohen's κ	0.563	5242

Table 25. Evaluation metrics of the 5-class classifier on the feature set comprising 2 respiratory features and 8 cardiac features from the PPG signal (support refers to 50 s-sequences).

	Precision	Recall	F1-Score	Support for Each Class [Sequences]
Wake	0.639 ± 0.015	0.458 ± 0.027	0.534 ± 0.019	1276
REM	0.348 ± 0.009	0.295 ± 0.022	0.319 ± 0.017	1397
N1	0.125 ± 0.056	0.051 ± 0.014	0.072 ± 0.011	79
N2	0.934 ± 0.021	0.668± 0.023	0.779 ± 0.025	2387
N3	0.047 ± 0.025	0.087 ± 0.011	0.061 ± 0.016	103
Macro avg	0.419 ± 0.032	0.312 ± 0.020	0.353 ± 0.024	5242
Weighted avg	0.676 ± 0.018	0.497 ± 0.016	0.572 ± 0.018	5242
Total accuracy	0.497 (i.e., 49.7 ± 0.67%)	Cohen's κ	0.460	5242

As shown in Tables 24 and 25, both the EOG-only and PPG-only feature sets achieved substantially lower performance than the EEG-based and multimodal configurations presented in the previous sections. Overall, the EOG-only configuration provided slightly better performance than the PPG-only configuration.

4.4. Energy and Latency Characterization of the Five-Class Sleep Staging Model

To complement the memory footprint analysis, we evaluated the computational efficiency of the proposed sleep staging framework in terms of pipeline latency, throughput, and measured energy consumption per inference. Physical CPU power consumption was measured directly at the hardware level using Intel's Running Average Power Limit (RAPL) Model-Specific Registers (MSR) accessed via the PyLibreHardwareMonitor interface on the test platform (Intel Core i5-1334U, 13th Generation). Before execution, baseline CPU idle power (P_{idle}) was recorded over a steady-state sampling window. During execution, CPU active power (P_{active}) was continuously sampled, allowing us to isolate the net dynamic power delta ($\Delta P = P_{active} - P_{idle}$) attributable strictly to model execution.

Inference benchmarking was conducted over $N = 500$ execution iterations using TensorFlow graph execution (tf.function) preceded by 20 warm-up runs to eliminate initial compilation overhead. Latency was tracked using high-resolution hardware timers (time.perf_counter()), recording mean latency alongside 95th (P95) and 99th (P99) percentile tail latencies. Throughput was computed as the inverse of mean inference latency. The dynamic energy consumed per sequence inference (E_{inf}) was calculated as

$$E_{inf} = \Delta P \times T_{inf} \times 1000 \ [\text{mJ}] \tag{12}$$

where ΔP is the dynamic power in W and T_{inf} is the mean inference latency in seconds. It should be noted that Equation (12) quantifies the dynamic energy associated only with the execution of the trained classification model. Specifically, the reported energy corresponds to the CPU dynamic power attributable to model inference, measured after subtracting the processor idle power. Thus, this metric does not represent the total energy consumption of a complete wearable sleep-monitoring system, which would additionally include biosignal acquisition, sensor operation, analog front-end circuitry, signal conditioning, buffering, feature extraction, memory accesses, and communication overhead.

Furthermore, it must be considered that the sleep staging pipeline currently operates in a post-acquisition batch processing modality; in particular, the sleep staging pipeline loads the set of features previously extracted on the biosignal recording over the full night, applies standard scaling, generates the 10-sub-epochs sequences, and scores them into five classes (Wake, REM, N1, N2, and N3). To provide a complete latency profile, timing is explicitly disaggregated into:

- Feature Preprocessing Time (T_{prep}): Tabular data loading, standard scaling, and temporal sequence framing (10-epoch timesteps).
- Inference Latency (T_{inf}): Direct model evaluation time per sequence.

These hardware measurements demonstrate that the trained sleep staging model has a low computational cost when executed on edge-computing hardware targets. However, the reported latency and energy measurements refer exclusively to software execution (model inference) on the target processor and should not be interpreted as the overall energy consumption of a complete wearable implementation. Table 26 reports the full latency disaggregation, throughput, measured dynamic power, and model inference energy characterization across the three feature configurations considered (EEG-only, EEG + EOG + PPG, and EOG + PPG).

Table 26. Computational efficiency and hardware benchmarking of the five-class sleep staging model across the three feature sets, reporting disaggregated preprocessing time (T_{prep}), inference latency (T_{inf}), throughput, CPU usage, and measured dynamic CPU energy consumption per inference.

Feature Set	T_{prep} [ms]	Mean T_{inf} [ms]	P95 T_{inf} [ms]	P99 T_{inf} [ms]	Throughput [Sequences/s]	CPU Usage [%]	Dynamic Energy per Inference [mJ]
EEG-only	3938.78	2.55	3.87	5.26	392.76	7.10	3.51
EEG + EOG + PPG	4866.72	4.02	9.49	15.27	248.80	15.80	10.90
EOG + PPG	2961.07	3.93	4.44	5.65	254.63	2.20	2.88

4.5. Comparison of Class-Balancing Strategies

To evaluate the impact of the adopted class-balancing strategy and investigate the individual and combined effects of data augmentation and class-weighted loss, four different training configurations were compared:

- No balancing, where the original class distribution was maintained without any balancing strategy.
- Data augmentation only, where synthetic samples were generated for the underrepresented classes while no class weights were applied during training.
- Class-weighted loss only, where the original dataset was used, and class imbalance was handled exclusively through weighted loss.
- Data augmentation and class-weighted loss, corresponding to the strategy adopted in this work.

All experiments were performed using the same training, validation, and testing protocol to ensure a fair comparison. The comparison reported in Table 27 refers to the five-class classification task using the complete feature set.

Table 27. Performance comparison of different class-balancing strategies.

Balancing Strategy	Total Accuracy	Precision	Recall	F1-Score
No balancing	0.876 ± 0.041	0.918 ± 0.020	0.876 ± 0.017	0.892 ± 0.021
Data augmentation only	0.870 ± 0.037	0.917 ± 0.016	0.870 ± 0.014	0.895 ± 0.019
Class-weighted only	0.862 ± 0.031	0.906 ± 0.010	0.862 ± 0.015	0.879 ± 0.013
Data augmentation + Class-weighted loss	0.882 ± 0.034	0.918 ± 0.006	0.882 ± 0.003	0.897 ± 0.005

As shown in Table 27, the four class-balancing strategies yielded comparable performance, indicating that the proposed model is generally robust to the adopted balancing approach. Nevertheless, the combination of data augmentation and class-weighted loss consistently achieved the best results across all evaluation metrics, obtaining the highest total accuracy (0.882), precision (0.918), recall (0.882), and F1-score (0.897). These results suggest that the combined use of data augmentation and class-weighted loss provides complementary benefits for addressing class imbalance. Although the improvements over the other balancing strategies are modest, the combined approach consistently produced the best overall performance and was therefore adopted for all subsequent experiments.

Finally, considering the best balancing strategy (data augmentation + class-weighted loss), different loss function formulations were tested (Table 28). The results of the ablation study indicate that the cross-entropy function integrated in the proposed classifiers performs slightly better than other tested loss functions.

Table 28. Ablation on loss function formulations.

Loss Function Formulation	Total Accuracy	Precision	Recall	F1-Score
Standard Cross-Entropy	0.882 ± 0.034	0.918 ± 0.006	0.882 ± 0.003	0.897 ± 0.005
Pure Focal Loss ($\gamma = 1.0$)	0.862 ± 0.011	0.921 ± 0.019	0.862 ± 0.013	0.887 ± 0.018
Pure Focal Loss ($\gamma = 2.0$)	0.865 ± 0.013	0.922 ± 0.024	0.865 ± 0.015	0.889 ± 0.021
Transition-aware Loss	0.873 ± 0.029	0.904 ± 0.021	0.873 ± 0.030	0.885 ± 0.025

5. Discussion

The increasing demand for portable, non-invasive sleep monitoring solutions has driven the development of lightweight systems that minimize hardware complexity without compromising diagnostic reliability. In this context, multimodal approaches combining physiological signals have shown significant potential for improving sleep staging performance. Deep Learning architectures specifically designed for sequential data, such as LSTM networks, are particularly well-suited to capture the temporal dynamics of heterogeneous biosignals, including EEG, EOG, and cardiorespiratory features derived from PPG signals. This research work proposes a two-stage DL framework that leverages multimodal feature integration for five-class sleep classification. By combining frontal EEG and EOG PCs with respiratory and cardiac features extracted from PPG signals, the proposed approach achieves accurate and robust sleep staging while maintaining a lightweight architecture, making it a potential candidate for future implementation in wearable and embedded systems. Table 29 summarizes the model's total accuracy and memory footprint, trained and tested across the three feature sets, evaluated on a test set composed of subjects independent from the training set.

Table 29. Comparative results for each feature set used for training and testing the 5-class model, including total model accuracy and total memory footprint, with performance evaluated on an independent subject-level test set.

Feature Set Used for Training and Testing	Total Accuracy	Used Memory
56 PCs EEG	87.1%	2.95 MB
56 EEG PCs, 8 EOG PCs, 2 features related to the respiratory component, and 8 features related to the cardiac component of the PPG signal	88.2%	3.14 MB
8 EOG PCs, 2 features related to the respiratory component, and 8 features related to the cardiac component of the PPG signal (no EEG features)	52.7%	2.56 MB

The main contribution of the paper lies in the systematic integration and comparative evaluation of established techniques within a unified lightweight pipeline tailored to minimal-configuration sleep staging. By evaluating three distinct multimodal feature sets under identical training and testing protocols, this work isolates the relative contribution of EEG, EOG, and PPG-derived features to the five-class sleep staging performance. Furthermore, the explicit emphasis on memory footprint and computational efficiency distinguishes this study from prior work that primarily focuses on classification accuracy without considering implementation-related aspects. The findings thus provide a practical benchmark for lightweight sleep staging approaches and suggest the potential suitability of the proposed framework for future implementation in wearable or resource-constrained systems.

Compared with recent EEG-based research reporting five-class sleep staging algorithms, ref. [22] achieves comparable accuracy (90.01% vs. 87.1%) and Cohen's κ (0.839 vs. 0.844) using a dataset with a similar number of subjects (DREAMS), whereas ref. [23] reports a lower accuracy (83.4% vs. 87.1%) and Cohen's κ (0.772 vs. 0.844) on the same dataset. Nevertheless, it must be considered that differences in validation protocols, channel configurations, and preprocessing pipelines preclude a direct head-to-head comparison. Two prior studies conducted on the same BOAS dataset adopted different modeling strategies: one achieved an overall validation accuracy of 95% using a CNN-based architecture [28] with six EEG derivations (reported in Table 1); the same authors also evaluated an LSTM model on the same dataset, obtaining a validation accuracy of 88%. It is important to emphasize that these comparisons remain indirect because of differences in channel count, validation strategy, and data partitioning protocols across studies. The BOAS study reporting 95% accuracy [28] used six EEG leads and reported validation-set rather than test-set performance; direct comparison with the present work's single-channel frontal setup and test-set evaluation should therefore be interpreted cautiously. Another recent study [29] achieved 87% accuracy, 82% macro-F1 score, and 0.77 Cohen's κ using an RF classifier on the BOAS dataset, extracting features from three EEG derivations (Table 1); the Cz-Fz derivation yielded the best performance. These performances are lower than those of our model in terms of accuracy and Cohen's κ (87.1% accuracy, 0.844 Cohen's κ), but comparable to the macro-F1 score (0.740%) on the same dataset using different EEG leads. However, in both studies, limited methodological details are provided regarding dataset construction, preprocessing, and feature extraction, making direct comparisons more challenging.

Despite its minimal frontal setup and reduced feature representation, the proposed two-stage LSTM-based framework achieved 87.1% total accuracy on the same dataset, falling within the performance range reported by the LSTM model in ref. [28] (88% validation accuracy) and the RF-based approach in ref. [29] (87% accuracy). In addition, compared to more computationally demanding architectures such as ResNet50-1D, which rely on six EEG derivations, the proposed framework offers a favorable trade-off between performance and model complexity while using only a single frontal channel. These findings indicate that the cascaded architecture shows potential for automatic sleep staging under minimal setup constraints while effectively controlling feature dimensionality. From a computational perspective, the proposed framework has a lightweight memory footprint. The WRN model requires 1.81 MB, and the NREM model 1.15 MB, for a total of 2.95 MB (Table 29) for the full cascaded architecture. This compact footprint suggests potential suitability for future deployment on resource-constrained embedded platforms.

Although using the single frontal lead limits the information from other brain regions (e.g., centro-parietal or occipital) that could aid sleep-stage discrimination, the two-stage LSTM-based framework achieved good five-stage sleep classification with a minimal electrode setup, demonstrating the effectiveness of frontal-temporal EEG features (F4-F3) and the feature selection strategy.

Compared with prior works reporting high accuracy with multimodal inputs, it can be observed that the proposed DL framework achieved performance within the range of published results using a comprehensive feature set (56 EEG PCs, eight EOG PCs, two respiratory, and eight cardiac PPG features). Prior studies have shown that combining EEG with other biosignals can improve sleep-stage classification; for instance, CNN-LSTM models using EEG and EOG achieved 83% accuracy and 0.78 Cohen's κ in ref. [30], both lower than the values obtained by our model (88.2% accuracy and 0.857 Cohen's κ). Also, the accuracy reported in ref. [31] (91%) was comparable to that achieved by the proposed model (88.2%). Furthermore, bagging ensembles with EEG, HRV, and ECG-derived respiration

reached 95% accuracy on clinical datasets [32]. Under the multimodal configuration, our two-stage LSTM-based framework achieved 88.2% five-class accuracy and a macro-average F1-score of 0.737. These findings should be interpreted with caution, as the comparisons drawn here are indirect. Differences across studies, including dataset composition, subject demographics, electrode montages, feature extraction pipelines, and validation strategies, make the reported performance figures not directly interchangeable. Within these limitations, the current results suggest that integrating neural, ocular, and cardiorespiratory information may enhance the separability of REM and transitional NREM stages, refine stage boundaries, and reduce misclassification between physiologically overlapping states, all while preserving a lightweight architecture suitable for embedded applications. Confirming these relative advantages would require head-to-head benchmarking under standardized protocols.

The complete multimodal feature set resulted in a modest improvement in overall classification accuracy, whereas the macro-average F1-score remained comparable to that of the EEG-only configuration. The observed gains were mainly associated with REM classification, while N1 performance remained largely unchanged. These results suggest that ocular and cardiorespiratory signals may provide complementary information to EEG in the proposed framework, supporting the development of robust, low-cost, and portable sleep-monitoring solutions.

In contrast, the feature set excluding EEG (EOG + PPG features only) yielded lower classification performance. Nevertheless, EOG and cardiorespiratory signals still provided useful complementary information, particularly for coarse stage discrimination. Without EEG, fine-grained classification suffers, reducing reliability for transitional and REM-related patterns. The lower precision in REM and N3, along with poor N1 detection, underscores the importance of cortical dynamics captured by EEG for automatic sleep staging. Despite the reduced feature set, the framework remains lightweight and may represent a suitable candidate for future embedded and wearable applications; indeed, the memory footprint of the cascaded architecture is 2.56 MB (WRN: 1.59 MB, NREM: 0.98 MB), a reduction of 0.58 MB compared to the model trained and tested with the complete feature set (Table 29). These results demonstrate a viable compromise between performance and resource efficiency when EEG signals are excluded.

The additional feature-group ablation analysis further clarifies the individual contribution of the non-EEG modalities. Both the EOG-only and PPG-only configurations showed substantially lower performance than the EEG-based and multimodal feature sets. Although the EOG-only configuration achieved slightly higher overall performance than the PPG-only configuration, the results indicate that neither modality alone provides sufficient information for reliable five-class sleep staging. Instead, their contribution appears to be complementary when combined with EEG features.

Furthermore, in Section 4.4, the characterization of the five-class model in terms of inference latency and energy expenditure per inference was reported. The benchmarking results show that the proposed model achieves consistently low inference latency across the three evaluated feature sets, with mean latency values of 2.55 ms for EEG-only, 4.02 ms for EEG + EOG + PPG, and 3.93 ms for EOG + PPG (Table 26). These values indicate that the model can support near-real-time processing of sleep epochs, an essential requirement for continuous sleep-monitoring applications. In particular, the EEG + EOG + PPG configuration exhibits higher preprocessing time, latency, and energy consumption than the other two feature sets, due to the increased dimensionality of the input data and the resulting computational overhead. Among the evaluated feature sets, the EOG + PPG (without EEG) feature set offers the best balance in terms of computational cost, achieving the lowest energy consumption per inference (2.88 mJ) and CPU utilization (2.20%), compared to

3.51 mJ and 7.10% for the EEG-only configuration and 10.90 mJ and 15.80% for the EEG + EOG + PPG configuration. This reduction in energy consumption and CPU utilization is mainly due to the lower number of input features/PCs, which decreases the computational burden of both preprocessing and inference. However, the EOG + PPG feature set produces the lowest classification accuracy among the three evaluated configurations due to the lack of EEG PCs. Overall, the results demonstrate that the proposed model, evaluated across all three feature sets, maintains stable, low-latency inference while reducing computational cost, suggesting its potential suitability for future deployment in embedded systems.

The proposed framework was evaluated using a subject-level data partitioning strategy, ensuring that subjects included in the test set were never seen during training. This protocol reduces the likelihood of data leakage arising from overlapping temporal windows and provides a more realistic assessment of the model's ability to generalize to unseen individuals. Furthermore, cross-dataset validation on the RichSleep (78.2%) and ISRUC (77.3%) datasets, acquired using different recording hardware, electrode configurations, and subject populations, further supports the generalization capability of the proposed framework beyond the BOAS dataset.

To further assess the generalization capability of the proposed framework, an additional Leave-One-Subject-Out (LOSO) cross-validation analysis was performed. In this protocol, all recordings from one subject were excluded from training and used exclusively for testing, while all preprocessing, feature selection, dimensionality reduction, data augmentation, and model training steps were repeated independently within each training fold. This evaluation provides a stringent subject-independent assessment and reduces the possibility of optimistic performance estimates caused by subject overlap between training and testing. Table 30 reports the mean and standard deviation of precision, recall, F1-score, and overall accuracy obtained using LOSO validation.

Table 30. Mean and standard deviation of precision, recall, F1-score, and total accuracy obtained using the LOSO validation.

	Precision	**Recall**	**F1-Score**
Macro avg	0.561 ± 0.084	0.597 ± 0.096	0.569 ± 0.097
Weighted avg	0.891 ± 0.028	0.865 ± 0.031	0.874 ± 0.030
Total accuracy	0.869 ± 0.029 (i.e., 86.9% ± 2.9%)		

As reported in Table 30, the LOSO evaluation yielded a total accuracy of 86.7% ± 3.1%, with weighted-average F1-score, precision, and recall values of 0.870 ± 0.032, 0.888 ± 0.030, and 0.862 ± 0.032, respectively. Although the macro-average metrics were lower (F1-score = 0.563 ± 0.101), reflecting the persistent difficulty of accurately classifying minority sleep stages (particularly N1 and N3), the overall performance remained consistent across the different LOSO folds. These results indicate that the proposed framework retains good subject-independent generalization capability despite inter-subject variability and support the robustness of the proposed methodology under a stringent validation protocol.

5.1. Validation on Public Datasets Different from the BOAS

To rigorously evaluate the generalization capability of our proposed two-stage LSTM framework beyond the BOAS dataset, we conducted external validation on two independent public sleep databases: the RichSleep [72] and the ISRUC-Sleep dataset [73]. These datasets were selected because their available signal modalities allowed the extraction of the main feature groups required by the proposed framework, including EEG, EOG, and cardiorespiratory features. Nevertheless, neither dataset provides the PPG signal; thus, the same cardiorespiratory features originally derived from PPG in the BOAS dataset

were extracted from the available ECG signal instead. Specifically, the cardiorespiratory feature definitions used on the BOAS dataset were adapted to extract features from the ECG signal. In detail, for the respiratory features, the respiratory component of the ECG was extracted by preprocessing the raw signal with a band-pass filter with 0.1 and 0.5 Hz cut-off frequencies, identifying the inspiration and expiration times using a peak detection algorithm, and applying the same definition in Table 5. For cardiac features, the raw ECG signal was preprocessed with a band-pass filter with 0.5 and 40 Hz cut-off frequencies; the PP interval was substituted with the interval between R peaks (RR interval), obtained through a QRS detection algorithm, enabling extraction of time-domain features reported in Table 6. The frequency-domain features maintain the same definition as in Table 6. Therefore, the feature representation was maintained unchanged at the physiological feature level, while the signal source used for feature extraction was modified due to dataset availability constraints. The extracted ECG-based features preserved the same units as the original PPG-derived features, enabling the application of the proposed framework without modifying the input feature set. This substitution was performed to enable external evaluation while maintaining the physiological information required by the framework. The RichSleep dataset comprises PSG recordings from 24 healthy subjects, including six EEG channels (F3, F4, C3, C4, O1, and O2), two EOG channels, and one ECG channel. The available EEG, EOG, and ECG signals were processed according to the same preprocessing pipeline adopted for the BOAS dataset, including signal preprocessing, segmentation into 5 s epochs, and extraction of the corresponding feature sets described in Section 3.2. Instead, the ISRUC-Sleep dataset consists of three subgroups: Subgroup 1 (SG1) includes 100 subjects with evidence of sleep disorders; Subgroup 2 (SG2) contains data from eight subjects, acquired across two acquisition sessions; and Subgroup 3 (SG3) is the control group, containing data from 10 healthy subjects. For the ISRUC dataset, the model was evaluated on SG2 and SG3 because these subsets provided the signal modalities required to extract the complete feature configuration adopted in this study, including EEG, EOG, and ECG-derived cardiorespiratory features obtained as described above. These subgroups were selected to provide an independent evaluation cohort composed of both healthy subjects and subjects with different acquisition sessions, while maintaining consistency with the feature configuration adopted in this study, including EEG, EOG, and ECG-derived cardiorespiratory features.

Using the same preprocessing and feature-extraction pipeline outlined in Section 3.2, we employed a zero-shot evaluation strategy in which the model was directly assessed on unseen target datasets without any additional training or adaptation. No information from the external datasets was used for model recalibration, feature selection, PCA fitting, threshold adjustment, or parameter optimization. The preprocessing parameters, including feature scaling and transformation steps, were inherited from the BOAS training procedure and applied unchanged to the external datasets. This protocol enabled the assessment of the model transferability and generalization capability under cross-dataset conditions, including evaluation on subjects with sleep disorders. Table 31 summarizes the performance (mean $\pm$ σ) of the proposed model across the two datasets considered.

Table 31. Validation results (mean$\pm$ σ) of the proposed two-stage LSTM model on RichSleep and ISRUC (SG2 + SG3) datasets.

Dataset	Accuracy	Precision	Recall	F1-Score
RichSleep	0.782 $\pm$ 0.049	0.801 $\pm$ 0.045	0.778 $\pm$ 0.042	0.759 $\pm$ 0.046
ISRUC (SG2 + SG3)	0.773 $\pm$ 0.038	0.778 $\pm$ 0.041	0.766 $\pm$ 0.042	0.763 $\pm$ 0.039

For the RichSleep dataset, the proposed model achieved a mean accuracy of 78.2% (σ = 4.9%) with a mean F1-score of 75.9% (σ = 4.6%). Similarly, the ISRUC dataset reached a mean accuracy of 77.3% and a mean F1-score of 76.3%. The lower cross-dataset performance compared with the BOAS evaluation indicates that the proposed framework can be transferred to independent datasets, although its performance remains affected by inter-dataset variability. These differences may be related to variations in recording hardware, sampling rates, electrode placement, preprocessing procedures, and population characteristics between BOAS and the target datasets (RichSleep and ISRUC). Previous studies have also reported reduced performance when sleep staging models are evaluated on external datasets, highlighting the difficulty of achieving complete robustness across heterogeneous acquisition conditions. In particular, in cross-dataset testing, well-known sleep staging models such as TinySleepNet and XSleepNet2 showed a significant performance drop compared to those achieved in the validation phase [74,75]. Therefore, these external validation results should be interpreted as evidence of partial generalization capability rather than complete robustness.

5.2. Comparative Analysis with Other Lightweight Deep Learning Architectures

Finally, an additional comparison was conducted with two representative lightweight Deep Learning architectures, TinySleepNet [76] and SeqSleepNet [68]. To ensure a fair comparison, both architectures were adapted to a two-stage cascaded classification strategy, employing two instances of the corresponding network to perform the first-stage (Wake/REM/NREM) and second-stage (N1/N2/N3) classifications. All models were trained using only the complete feature set and the same training and evaluation protocol. Table 32 reports the weighted precision, weighted recall, weighted F1-score, and total accuracy achieved by the three approaches.

Table 32. Performance comparison between the proposed cascaded LSTM framework and the cascaded TinySleepNet and SeqSleepNet implementations using the same training and evaluation protocol on the BOAS dataset.

Model	Total Accuracy	Precision	Recall	F1-Score
Our model	0.882	0.918	0.882	0.897
TinySleepNet [76]	0.871	0.901	0.871	0.891
SeqSleepNet [68]	0.845	0.673	0.845	0.861

As reported in Table 32, the proposed framework achieved the highest performance among the three evaluated approaches, slightly outperforming the cascaded TinySleepNet implementation and more clearly surpassing the cascaded SeqSleepNet model across the considered metrics. These results suggest that the proposed two-stage strategy can provide competitive performance while maintaining a lightweight architecture.

6. Conclusions

This study presents a two-stage LSTM-based framework for five-stage sleep classification, systematically trained and tested across three feature sets (EEG-only, multimodal-EEG + EOG + PPG, EOG + PPG) extracted from the BOAS database. Using the EEG-only feature set, the cascaded model achieves good performance, reaching 87.1% total accuracy and a macro-average F1-score of 0.740, while maintaining a compact memory footprint of 2.95 MB, confirming that a minimal frontal setup (one EEG, one EOG, and one PPG) can deliver reliable sleep staging. Using the complete multimodal feature set improves performance to 88.2% total accuracy and a macro-average F1-score of 0.737, with a modest increase in

memory usage to 3.14 MB, suggesting that complementary ocular and cardiorespiratory information enhances stage separability while maintaining low memory requirements.

Nevertheless, using the EEG-excluded feature set drastically reduces the framework's performance, resulting in 52.7% total accuracy and an F1-score of 0.382. Overall, these findings confirm that EEG features serve as the primary discriminative backbone of the system, while EOG and PPG features provide complementary information that refines classification decisions.

The benchmarking analysis demonstrates that the proposed model features low inference latency (2.55–4.02 ms) and energy per inference (2.88–10.9 mJ per inference) across the three feature sets, with the EOG + PPG configuration offering the lowest computational cost at the expense of reduced accuracy, highlighting a clear trade-off between performance and latency. These results suggest that the proposed framework offers a favorable accuracy-efficiency balance for future implementation in embedded and wearable sleep-monitoring systems, although this was not evaluated in the present study. Finally, the proposed model was extensively validated. In particular, LOSO cross-validation on the BOAS database yielded 86.9% $\pm$ 2.9% accuracy, while cross-dataset validation on two public datasets (RichSleep and ISRUC) provided 78.2% and 77.3% accuracy, respectively. As expected given the domain shift between datasets, the results show a performance drop of -10.0% and -10.9% relative to the test phase, reflecting differences in recording hardware, electrode placement, and population characteristics compared to the BOAS dataset. Comparative analysis against cascaded TinySleepNet (87.1%) and SeqSleepNet (84.5%) shows the proposed model's competitive performance.

Future work will focus on deploying the proposed framework on low-power wearable hardware platforms to further characterize memory usage and energy consumption under real-world operating conditions. Furthermore, future developments will include integrating dedicated preprocessing techniques for biosignal artifact mitigation, such as independent component analysis (ICA), to evaluate their impact on EEG, EOG, and PPG signal quality and model robustness under motion-contaminated wearable scenarios. In addition, future research will explore the adoption of a multi-resolution input window strategy to capture temporal sleep dynamics at different time scales and further improve sleep staging performance.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/app16168164/s1. File S1: Script_ Sleep_Staging_Model_Train_Test.py.

Author Contributions: Conceptualization, R.D.F., M.P., C.D.-V.-S. and P.V.; methodology, R.D.F., C.D.-V.-S. and R.V.; validation, R.D.F., R.V. and P.V.; formal analysis, R.D.F. and M.P.; investigation, R.D.F. and M.P.; software, R.D.F. and M.P.; resources, C.D.-V.-S., P.V. and R.V.; data curation, R.D.F., M.P. and P.V.; writing—original draft preparation, R.D.F., M.P. and P.V.; writing—review and editing, R.D.F., C.D.-V.-S., R.V. and P.V.; visualization, R.D.F. and C.D.-V.-S.; supervision, R.V. and P.V.; funding acquisition, P.V. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Italian Space Agency, medium-term research and development projects related to the "ASI Academic Space Research Days", project "PoliSOnnigrafo Multi-sensore Non-Invasivo Indossabile per Astronauti con MONITOraggio Remoto dei parametri vitali e della qualità del sonno-SOMNIIA MONITOR" (CUP F33C23000260005).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DL	Deep Learning
LSTM	Long Short-Term Memory
EEG	Electroencephalogram
EOG	Electrooculogram
PPG	Photoplethysmogram
mRMR	Minimum Redundancy Maximum Relevance
PCA	Principal Component Analysis
PSG	Polysomnography
ASI	Italian Space Agency
ML	Machine Learning
SEMs	Slow Eye Movements
BR	Breathing Rate
REM	Rapid Eye Movement
NREM	Non-Rapid Eye Movement
IER	Inspiratory-to-Expiratory Ratio
HR	Heart Rate
HRV	Heart Rate Variability
ECG	Electrocardiogram
SDNN	Standard Deviation of NN intervals
RMSSD	Root Mean Square of Successive Differences
LF	Low Frequency
HF	High Frequency
BOAS	Bitbrain Open Access Sleep
TCN	Temporal Convolutional Network
AASM	American Academy of Sleep Medicine
R&K	Rechtschaffen and Kales
EEMD	Ensemble Empirical Mode Decomposition
SHHS	Sleep Heart Health Study
NSP	Null Space Pursuit
RVMs	Relevance Vector Machines
SVM	Support Vector Machine
RF	Random Forest
CNN	Convolutional Neural Network
EDR	ECG-Derived Respiration
MLP	Multilayer Perceptron
MGH	Massachusetts General Hospital
EMG	Electromyography
RMS	Root Mean Square
MI	Mutual Information
MIQ	Mutual Information Quotient
ReLU	Rectified Linear Unit
ZCR	Zero-Crossing Rate
AAC	Average Amplitude Change
IQR	Interquartile Range
SSI	Simple Square Integral
RSP	Relative Spectral Power
SWI	Slow Wave Index
SEFd	Spectral Edge Frequency
AP	Absolute Power
SVD	Singular Value Decomposition
LZC	Lempel–Ziv Complexity

SD1	Standard deviation of consecutive PP interval differences
SD2	Long-term deviation of the Poincaré representation
CPU	Central Processing Unit
TDP	Thermal Design Power
ISRUC	Institute of Systems and Robotics, University of Coimbra
ICA	Independent Component Analysis

References

1. Hirshkowitz, M. Normal Human Sleep: An Overview. *Med. Clin. N. Am.* **2004**, *88*, 551–565. [CrossRef] [PubMed]
2. Olbrich, E.; Achermann, P.; Meier, P.F. Dynamics of Human Sleep EEG. *Neurocomputing* **2003**, *52–54*, 857–862. [CrossRef]
3. Worley, S.L. The Extraordinary Importance of Sleep: The Detrimental Effects of Inadequate Sleep on Health and Public Safety Drive an Explosion of Sleep Research. *P T Peer-Rev. J. Formul. Manag.* **2018**, *43*, 758–763.
4. Gerstenslager, B.; Slowik, J.M. Sleep Study. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2025.
5. Kwon, S.; Kim, H.; Yeo, W.-H. Recent Advances in Wearable Sensors and Portable Electronics for Sleep Monitoring. *iScience* **2021**, *24*, 102461. [CrossRef] [PubMed]
6. Eldele, E.; Chen, Z.; Liu, C.; Wu, M.; Kwoh, C.-K.; 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. [CrossRef] [PubMed]
7. De Fazio, R.; Mattei, V.; Al-Naami, B.; De Vittorio, M.; Visconti, P. Methodologies and Wearable Devices to Monitor Biophysical Parameters Related to Sleep Dysfunctions: An Overview. *Micromachines* **2022**, *13*, 1335. [CrossRef] [PubMed]
8. De Fazio, R.; Mastronardi, V.M.; De Vittorio, M.; Spongano, L.; Fachechi, L.; Rizzi, F.; Visconti, P. A Sensorized Face Mask to Monitor Sleep and Health of the Astronauts: Architecture Definition, Sensing Section Development and Biosignals' Acquisition. In Proceedings of the 2024 9th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 25–28 June 2024; pp. 1–6. [CrossRef]
9. Buzsáki, G. *Rhythms of the Brain*; Oxford University Press: Oxford, UK, 2006; ISBN 978-0-19-530106-9.
10. Caffier, P.P.; Erdmann, U.; Ullsperger, P. Experimental Evaluation of Eye-Blink Parameters as a Drowsiness Measure. *Eur. J. Appl. Physiol.* **2003**, *89*, 319–325. [CrossRef] [PubMed]
11. Pizza, F.; Fabbri, M.; Magosso, E.; Ursino, M.; Provini, F.; Ferri, R.; Montagna, P. Slow Eye Movements Distribution during Nocturnal Sleep. *Clin. Neurophysiol.* **2011**, *122*, 1556–1561. [CrossRef] [PubMed]
12. Yang, J.; Keller, J.M.; Popescu, M.; Skubic, M. Sleep stage recognition using respiration signal. In Proceedings of the 2016 38th International Conference of IEEE Engineering in Medicine & Biology Society, Orlando, FL, USA, 16–20 August 2016; pp. 2843–2846. [CrossRef] [PubMed]
13. Yuda, E.; Yoshida, Y. Synchronization of Inhalation/Exhalation Ratio and Heart Rate Variability During Spontaneous Breathing. *Electronics* **2025**, *14*, 1903. [CrossRef]
14. Campana, L.M.; Owens, R.; Butler, J.; Suki, B.; Malhotra, A. Variability of Respiratory Mechanics during Sleep in Overweight and Obese Subjects with and without Asthma. *Respir. Physiol. Neurobiol.* **2013**, *186*, 290–295. [CrossRef] [PubMed]
15. Lyu, J.; Shi, W.; Zhang, C.; Yeh, C.-H. A Novel Sleep Staging Method Based on EEG and ECG Multimodal Features Combination. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 4073–4084. [CrossRef] [PubMed]
16. Simões, H.; Pires, G.; Nunes, U.; Silva, V. Feature Extraction and Selection for Automatic Sleep Staging Using EEG. In Proceedings of the 7th International Conference on Informatics in Control, Automation and Robotics, Madeira, Portugal, 15–18 June 2010; pp. 128–133. [CrossRef]
17. 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. [CrossRef] [PubMed]
18. Leino, A.; Korkalainen, H.; Kalevo, L.; Nikkonen, S.; Kainulainen, S.; Ryan, A.; Duce, B.; Sipila, K.; Ahlberg, J.; Sahlman, J.; et al. Deep Learning Enables Accurate Automatic Sleep Staging Based on Ambulatory Forehead EEG. *IEEE Access* **2022**, *10*, 26554–26566. [CrossRef]
19. Khalili, E.; Mohammadzadeh Asl, B. Automatic Sleep Stage Classification Using Temporal Convolutional Neural Network and New Data Augmentation Technique from Raw Single-Channel EEG. *Comput. Methods Programs Biomed.* **2021**, *204*, 106063. [CrossRef] [PubMed]
20. Zhao, S.; Long, F.; Wei, X.; Ni, X.; Wang, H.; Wei, B. Evaluation of a Single-Channel EEG-Based Sleep Staging Algorithm. *Int. J. Environ. Res. Public Health* **2022**, *19*, 2845. [CrossRef] [PubMed]
21. Ghimatgar, H.; Kazemi, K.; Helfroush, M.S.; Aarabi, A. An Automatic Single-Channel EEG-Based Sleep Stage Scoring Method Based on Hidden Markov Model. *J. Neurosci. Methods* **2019**, *324*, 108320. [CrossRef] [PubMed]
22. Liu, C.; Tan, B.; Fu, M.; Li, J.; Wang, J.; Hou, F.; Yang, A. Automatic Sleep Staging with a Single-Channel EEG Based on Ensemble Empirical Mode Decomposition. *Phys. Stat. Mech. Its Appl.* **2021**, *567*, 125685. [CrossRef]

23. Xiao, W.; Linghu, R.; Li, H.; Hou, F. Automatic Sleep Staging Based on Single-Channel EEG Signal Using Null Space Pursuit Decomposition Algorithm. *Axioms* **2023**, *12*, 30. [CrossRef]
24. Imtiaz, S.A.; Rodriguez-Villegas, E. A Low Computational Cost Algorithm for REM Sleep Detection Using Single Channel EEG. *Ann. Biomed. Eng.* **2014**, *42*, 2344–2359. [CrossRef] [PubMed]
25. Huang, C.-S.; Lin, C.-L.; Ko, L.-W.; Liu, S.-Y.; Su, T.-P.; Lin, C.-T. Knowledge-Based Identification of Sleep Stages Based on Two Forehead Electroencephalogram Channels. *Front. Neurosci.* **2014**, *8*, 263. [CrossRef] [PubMed]
26. Huang, C.-S.; Lin, C.-L.; Ko, L.-W.; Liu, S.-Y.; Sua, T.-P.; Lin, C.-T. A hierarchical classification system for sleep stage scoring via forehead EEG signals. In Proceedings of the IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), Singapore, 16–19 April 2013. [CrossRef]
27. Popovic, D.; Khoo, M.; Westbrook, P. Automatic Scoring of Sleep Stages and Cortical Arousals Using Two Electrodes on the Forehead: Validation in Healthy Adults. *J. Sleep Res.* **2014**, *23*, 211–221. [CrossRef] [PubMed]
28. Duc, N.P. A comparative analysis of Deep Learning models for EEG-based sleep stage classification. In Proceedings of the International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC), Seoul, Republic of Korea, 7–10 July 2025. [CrossRef]
29. Yilmaz, G.; Martinez, C.S.; Braun, F.; Gnarra, O.; Seiler, A.; Schindler, K.; Schmidt, M.H.; Lemay, M.; Jorge, J. Sleep Staging for Wearable Electroencephalography Leveraging Machine Learning and Conventional Polysomnography Datasets. In Proceedings of the 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Copenhagen, Denmark, 14–18 July 2025; pp. 1–7. [CrossRef] [PubMed]
30. Korkalainen, H.; Leppanen, T.; Aakko, J.; Nikkonen, S.; Kainulainen, S.; Leino, A.; Duce, B.; Afara, I.O.; Myllymaa, S.; Toyras, J. Accurate Deep Learning-Based Sleep Staging in a Clinical Population with Suspected Obstructive Sleep Apnea. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2073–2081. [CrossRef] [PubMed]
31. Yildirim, O.; Baloglu, U.B.; Acharya, U.R. A Deep Learning Model for Automated Sleep Stages Classification Using PSG Signals. *Int. J. Environ. Res. Public Health* **2019**, *16*, 599. [CrossRef] [PubMed]
32. Rashidi, S.; Asl, B.M. Strength of Ensemble Learning in Automatic Sleep Stages Classification Using Single-Channel EEG and ECG Signals. *Med. Biol. Eng. Comput.* **2024**, *62*, 997–1015. [CrossRef] [PubMed]
33. Sun, H.; Ganglberger, W.; Panneerselvam, E.; Leone, M.J.; A Quadri, S.; Goparaju, B.; A Tesh, R.; Akeju, O.; Thomas, R.J.; Westover, M.B. Sleep Staging from Electrocardiography and Respiration with Deep Learning. *Sleep* **2020**, *43*, zsz306. [CrossRef] [PubMed]
34. Tăutan, A.-M.; Rossi, A.C.; de Francisco, R.; Ionescu, B. Automatic sleep stage detection: A study on the influence of various PSG input Signals. In Proceedings of the 2020 Annual International Conference of the IEEE Engineering in Medicine & Biology Society, Montreal, QC, Canada, 20–24 July 2020; pp. 5330–5334. [CrossRef] [PubMed]
35. Berry, R.B.; Albertario, C.L.; Harding, S.M. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications: Version 2.5*; American Academy of Sleep Medicine: Darien, IL, USA, 2018.
36. Danker-Hopfe, H.; Anderer, P.; Zeitlhofer, J.; Boeck, M.; Dorn, H.; Gruber, G.; Heller, E.; Loretz, E.; Moser, D.; Parapatics, S.; et al. Interrater Reliability for Sleep Scoring According to the Rechtschaffen & Kales and the New AASM Standard. *J. Sleep Res.* **2009**, *18*, 74–84. [CrossRef] [PubMed]
37. Rosenberg, R.S.; Van, H.S. The American Academy of Sleep Medicine Inter-Scorer Reliability Program: Sleep Stage Scoring. *J. Clin. Sleep Med.* **2013**, *9*, 81–87. [CrossRef] [PubMed]
38. Esparza-Iaizzo, M.; Sierra-Torralba, M.; Klinzing, J.G.; Minguez, J.; Montesano, L.; López-Larraz, E. Automatic Sleep Scoring for Real-Time Monitoring and Stimulation in Individuals with and without Sleep Apnea. *Comput. Biol. Med.* **2026**, *205*, 111560. [CrossRef] [PubMed]
39. Metzner, C.; Schilling, A.; Traxdorf, M.; Schulze, H.; Krauss, P. Sleep as a Random Walk: A Super-Statistical Analysis of EEG Data across Sleep Stages. *Commun. Biol.* **2021**, *4*, 1385. [CrossRef] [PubMed]
40. Depoortere, H.; Francon, D.; Granger, P.; Terzano, M.G. Evaluation of the Stability and Quality of Sleep Using Hjorth's Descriptors. *Physiol. Behav.* **1993**, *54*, 785–793. [CrossRef] [PubMed]
41. Marino, S.; Silveri, G.; Bonanno, L.; De Salvo, S.; Cartella, E.; Miladinović, A.; Ajčević, M.; Accardo, A. Linear and Non-Linear Analysis of EEG During Sleep Deprivation in Subjects with and Without Epilepsy. In *XV Mediterranean Conference on Medical and Biological Engineering and Computing—MEDICON 2019*; Henriques, J., Neves, N., de Carvalho, P., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 125–132. [CrossRef]
42. Taran, S.; Sharma, P.C.; Bajaj, V. Automatic Sleep Stages Classification Using Optimize Flexible Analytic Wavelet Transform. *Knowl.-Based Syst.* **2020**, *192*, 105367. [CrossRef]
43. Aamodt, A.; Nilsen, A.S.; Markhus, R.; Kusztor, A.; HasanzadehMoghadam, F.; Kauppi, N.; Thürer, B.; Storm, J.F.; Juel, B.E. EEG Lempel-Ziv Complexity Varies with Sleep Stage, but Does Not Seem to Track Dream Experience. *Front. Hum. Neurosci.* **2023**, *16*, 987714. [CrossRef] [PubMed]
44. Li, H.; Peng, C.; Ye, D. A Study of Sleep Staging Based on a Sample Entropy Analysis of Electroencephalogram. *Biomed. Mater. Eng.* **2015**, *26*, S1149–S1156. [CrossRef] [PubMed]

45. Aboalayon, K.A.I.; Ocbagabir, H.T.; Faezipour, M. Efficient sleep stage classification based on EEG signals. In Proceedings of the IEEE Long Island Systems, Applications and Technology Conference (LISAT), Farmingdale, NY, USA, 2 May 2014. [CrossRef]
46. Jan Dijk, D.; Beersma, D.G.M.; Daan, S.; Bloem, G.M.; Hoofdakker, R.H.V.D. Quantitative Analysis of the Effects of Slow Wave Sleep Deprivation during the First 3 h of Sleep on Subsequent EEG Power Density. *Eur. Arch. Psychiatry Neurol. Sci.* **1987**, *236*, 323–328. [CrossRef] [PubMed]
47. Wu, J.; Zhou, Q.; Li, J.; Chen, Y.; Shao, S.; Xiao, Y. Decreased Resting-State Alpha-Band Activation and Functional Connectivity after Sleep Deprivation. *Sci. Rep.* **2021**, *11*, 484. [CrossRef] [PubMed]
48. Gorlova, S.; Ichiba, T.; Nishimaru, H.; Takamura, Y.; Matsumoto, J.; Hori, E.; Nagashima, Y.; Tatsuse, T.; Ono, T.; Nishijo, H. Non-Restorative Sleep Caused by Autonomic and Electroencephalography Parameter Dysfunction Leads to Subjective Fatigue at Wake Time in Shift Workers. *Front. Neurol.* **2019**, *10*, 66. [CrossRef] [PubMed]
49. Krystal, A.D.; Edinger, J.D.; Wohlgemuth, W.K.; Marsh, G.R. NREM Sleep EEG Frequency Spectral Correlates of Sleep Complaints in Primary Insomnia Subtypes. *Sleep* **2002**, *25*, 626–636. [CrossRef]
50. Peng, H.; Hu, B.; Zheng, F.; Fan, D.; Zhao, W.; Chen, X.; Yang, Y.; Cai, Q. A Method of Identifying Chronic Stress by EEG. *Pers. Ubiquitous Comput.* **2013**, *17*, 1341–1347. [CrossRef]
51. Hussain, I.; Hossain, M.A.; Jany, R.; Bari, M.A.; Uddin, M.; Kamal, A.R.M.; Ku, Y.; Kim, J.-S. Quantitative Evaluation of EEG-Biomarkers for Prediction of Sleep Stages. *Sensors* **2022**, *22*, 3079. [CrossRef] [PubMed]
52. Carli, F.D.; Nobili, L.; Beelke, M.; Watanabe, T.; Smerieri, A.; Parrino, L.; Terzano, M.G.; Ferrillo, F. Quantitative Analysis of Sleep EEG Microstructure in the Time–Frequency Domain. *Brain Res. Bull.* **2004**, *63*, 399–405. [CrossRef] [PubMed]
53. Jobert, M.; Schulz, H.; Jähnig, P.; Tismer, C.; Bes, F.; Escola, H. A Computerized Method for Detecting Episodes of Wakefulness During Sleep Based on the Alpha Slow-Wave Index (ASI). *Sleep* **1994**, *17*, 37–46. [CrossRef]
54. You, Y.; Zhong, X.; Liu, G.; Yang, Z. Automatic Sleep Stage Classification: A Light and Efficient Deep Neural Network Model Based on Time, Frequency and Fractional Fourier Transform Domain Features. *Artif. Intell. Med.* **2022**, *127*, 102279. [CrossRef] [PubMed]
55. Van Hese, P.; Philips, W.; De Koninck, J.; Van de Walle, R.; Lemahieu, I. Automatic Detection of Sleep Stages Using the EEG. In Proceedings of the 23rd Annual International Conference of the IEEE Engin. Medicine and Biology Society, Istanbul, Turkey, 25–28 October 2001; Volume 2, pp. 1944–1947. [CrossRef]
56. Ferreira, C.; Deslandes, A.; Moraes, H.; Cagy, M.; Pompeu, F.; Basile, L.F.; Piedade, R.; Ribeiro, P. Electroencephalographic Changes after One Nigth of Sleep Deprivation. *Arq. Neuropsiquiatr.* **2006**, *64*, 388–393. [CrossRef] [PubMed]
57. Kozhemiako, N.; Mylonas, D.; Pan, J.Q.; Prerau, M.J.; Redline, S.; Purcell, S.M. Sources of Variation in the Spectral Slope of the Sleep EEG. *eNeuro* **2022**, *9*, 1–20. [CrossRef] [PubMed]
58. Michielli, N.; Acharya, U.R.; Molinari, F. Cascaded LSTM Recurrent Neural Network for Automated Sleep Stage Classification Using Single-Channel EEG Signals. *Comput. Biol. Med.* **2019**, *106*, 71–81. [CrossRef] [PubMed]
59. Fell, J.; Röschke, J.; Mann, K.; Schäffner, C. Discrimination of Sleep Stages: A Comparison between Spectral and Nonlinear EEG Measures. *Electroencephalogr. Clin. Neurophysiol.* **1996**, *98*, 401–410. [CrossRef] [PubMed]
60. Ma, Y.; Shi, W.; Peng, C.-K.; Yang, A.C. Nonlinear Dynamical Analysis of Sleep Electroencephalography Using Fractal and Entropy Approaches. *Sleep Med. Rev.* **2018**, *37*, 85–93. [CrossRef] [PubMed]
61. Banerjee, M.; Pal, N.R. Feature Selection with SVD Entropy: Some Modification and Extension. *Inf. Sci.* **2014**, *264*, 118–134. [CrossRef]
62. Lempel, A.; Ziv, J. On the Complexity of Finite Sequences. *IEEE Trans. Inf. Theory* **1976**, *22*, 75–81. [CrossRef]
63. De Gennaro, L.; Ferrara, M.; Ferlazzo, F.; Bertini, M. Slow Eye Movements and EEG Power Spectra during Wake-Sleep Transition. *Clin. Neurophysiol.* **2000**, *111*, 2107–2115. [CrossRef] [PubMed]
64. Stein, P.K.; Pu, Y. Heart Rate Variability, Sleep and Sleep Disorders. *Sleep Med. Rev.* **2012**, *16*, 47–66. [CrossRef] [PubMed]
65. Darbellay, G.A.; Vajda, I. Estimation of the Information by an Adaptive Partitioning of the Observation Space. *IEEE Trans. Inf. Theory* **1999**, *45*, 1315–1321. [CrossRef]
66. Ding, C.; Peng, H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *J. Bioinform. Comput. Biol.* **2005**, *3*, 185–205. [CrossRef] [PubMed]
67. 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*, 1998–2008. [CrossRef] [PubMed]
68. Phan, H.; Andreotti, F.; Cooray, N.; Chen, O.Y.; 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*, 400–410. [CrossRef] [PubMed]
69. Phan, H.; Mikkelsen, K.; Chen, O.Y.; Koch, P.; Mertins, A.; De Vos, M. SleepTransformer: Automatic Sleep Staging With Interpretability and Uncertainty Quantification. *IEEE Trans. Biomed. Eng.* **2022**, *69*, 2456–2467. [CrossRef] [PubMed]
70. He, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [CrossRef]
71. Buda, M.; Maki, A.; Mazurowski, M.A. A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef] [PubMed]

72. Liu, Z.; Zhang, Q.; Luo, S.; Qin, M.; Lei, X. RichSleep 2023. Available online: https://data.mendeley.com/datasets/35dp3db8yj/1 (accessed on 2 March 2026). [CrossRef]
73. Khalighi, S.; Sousa, T.; Santos, J.M.; Nunes, U. ISRUC-Sleep: A Comprehensive Public Dataset for Sleep Researchers. *Comput. Methods Programs Biomed.* **2016**, *124*, 180–192. [CrossRef] [PubMed]
74. Mai, X.; Song, B.; Luo, M.; Zhu, J.; Jiang, X.; Ma, X.; Lin, F.; Hu, X.; Peng, H.; Zhang, L.; et al. AISleep: Automated and Interpretable Sleep Staging from Single-Channel EEG Data. *Patterns* **2025**, *6*, 101367. [CrossRef] [PubMed]
75. Baumert, M.; Hartmann, S.; Phan, H. Automatic Sleep Staging for the Young and the Old—Evaluating Age Bias in Deep Learning. *Sleep Med.* **2023**, *107*, 18–25. [CrossRef] [PubMed]
76. 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)*; IEEE: New York, NY, USA, 2020; pp. 641–644. [CrossRef] [PubMed]