Article

MLS-Net: An Automatic Sleep Stage Classifier Utilizing Multimodal Physiological Signals in Mice

Chengyong Jiang [ID], Wenbin Xie, Jiadong Zheng, Biao Yan [ID], Junwen Luo * and Jiayi Zhang *[ID]

State Key Laboratory of Medical Neurobiology, MOE Frontiers Center for Brain Science, Institutes of Brain Science, Institute for Medical and Engineering Innovation, Department of Ophthalmology and Vision Science, Eye & ENT Hospital, Fudan University, Shanghai 200032, China; 22111520017@m.fudan.edu.cn (C.J.); xiewenbin80@163.com (W.X.); 20301050022@fudan.edu.cn (J.Z.); biaoyan@fudan.edu.cn (B.Y.)
* Correspondence: luojunwen@fudan.edu.cn (J.L.); jiayizhang@fudan.edu.cn (J.Z.)

Abstract: Over the past decades, feature-based statistical machine learning and deep neural networks have been extensively utilized for automatic sleep stage classification (ASSC). Feature-based approaches offer clear insights into sleep characteristics and require low computational power but often fail to capture the spatial–temporal context of the data. In contrast, deep neural networks can process raw sleep signals directly and deliver superior performance. However, their overfitting, inconsistent accuracy, and computational cost were the primary drawbacks that limited their end-user acceptance. To address these challenges, we developed a novel neural network model, MLS-Net, which integrates the strengths of neural networks and feature extraction for automated sleep staging in mice. MLS-Net leverages temporal and spectral features from multimodal signals, such as EEG, EMG, and eye movements (EMs), as inputs and incorporates a bidirectional Long Short-Term Memory (bi-LSTM) to effectively capture the spatial–temporal nonlinear characteristics inherent in sleep signals. Our studies demonstrate that MLS-Net achieves an overall classification accuracy of 90.4% and REM state precision of 91.1%, sensitivity of 84.7%, and an F1-Score of 87.5% in mice, outperforming other neural network and feature-based algorithms in our multimodal dataset.

Keywords: MLS-Net; ASSC; eye movements; multimodal signal

1. Introduction

Sleep and wakefulness are self-regulatory states of the body that are closely associated with numerous physiological processes. These processes include cognition, learning and memory, brain homeostasis, and various diseases [1,2]. Sleep disorders, such as sleep apnea [3], insomnia [4,5], cataplexy, and narcolepsy [6,7], represent one of the serious problems of modern society. Animal experiments have provided invaluable insights into the mechanisms underlying sleep regulation and sleep disorders [5,8,9]. Given the substantial reliance on rodent models in sleep research, developing rodent ASSC capable of discerning wake–sleep states for rodents would be greatly helpful for the diagnosis of disease and practice in preclinical sleep research [8]. However, rodent ASSC faces two major challenges: first, there is an urgent need to explore an appropriate model that can achieve high performance; second, the impact of multi-modal signal integration on the performance of automatic scoring algorithms is unclear.

The approaches applied to rodent ASSC can be delineated into two principal modalities [10]. The one is a conventional feature-based statistical machine learning approach, while the other involves a deep neural network. Feature-based statistical machine learning feeds the pre-defined feature vectors into vector-based classifiers such as support vector machine (SVM) [11,12], random forest (RF) [13], linear regression (LR), eXtreme Gradient Boosting (XGBoost) [14], linear discriminant analysis (LDA) [15], and so on. Deep neural networks process raw sleep biosignals directly, feeding them into well-designed neural networks for feature extraction and stage classification. This approach has gained popularity

due to its outstanding performance [16]. Despite the widespread application of deep neural networks, they present significant limitations. First, their generalization capabilities are often inadequate, particularly when applied to datasets involving animals from different experimental studies [17]. Second, the inherent complexity of these networks can result in the unintended capture of noise, which exacerbates overfitting [18,19]. This issue not only reduces the accuracy of the models but also impedes their practical application in real-time sleep scoring and sleep intervention [20]. This results in significant variability in accuracy, with reported performance ranging from 71.0% to 92.2% [21]. In contrast, feature-based approaches enable users to have clear information about the sleep characteristics underpinning the scoring process with a low energy cost [21]. However, they typically analyze each segment of data independently based on the pre-defined features and prior knowledge, without considering the spatial–temporal context. This limitation prevents them from extracting deeply concealed nonlinear characteristics from the signals, leading to less accurate classification [22].

Currently, rodent ASSC primarily relies on EEG and/or EMG signals due to the small size of rodents [23], which limits the acquisition and application of multi-modal physiological signals. Although alternative studies in humans have demonstrated the benefits of incorporating multi-modal signals [21] such as electromyogram (EOG), electrocardiogram (ECG) [24], respiration, snoring [25], and oxygen saturation [26] for improved classification performance [20,27], this approach has not been widely adopted in rodent research. Eye movements (EMs), as a fundamental indicator in distinguishing REM from NREM sleep [28], have significant potential to enhance the performance of ASSC in rodents, yet their application remains rare. Moreover, despite the extensive use of rodents like mice in preclinical sleep studies, there is a notable lack of publicly available multi-modal physiological datasets for rodents. Unlike human ASSC, which benefits from numerous open datasets for model validation, existing mouse sleep datasets, such as AccuSleep [29], UTSN-L [30], and SPINDLE [17], are limited to EEG/EMG data. The absence of a common multi-modal dataset contributes to discrepancies in published results across different research teams [23], highlighting the critical need for establishing an open multi-modal physiological dataset for mice to accelerate integrative physiology as well as behavioral studies.

To solve these dilemmas, we designed a neural network, MLS-Net, that combines the strengths of neural networks and feature extraction based on multi-model physiological signals for automated sleep staging in mice (Figure 1). The core innovation of MLS-Net lies in: (1) a tailor-designed bi-LSTM's proven capability in handling sequences and its effectiveness in capturing the spatial–temporal nonlinear characteristics inherent in consecutive sleep states during sleep; and (2) employs time–domain and frequency–domain expert features of multimodal signals, especially EMs, as inputs, which provides deterministic expert knowledge. This design can effectively address the issue of traditional deep neural networks experiencing a marked decline in classification accuracy for specific sleep stages, such as the REM stage [17]. Our studies, conducted on a robust dataset, show that MLS-Net achieves an impressive overall accuracy of 90.4% among other neural networks and feature-based algorithms. Notably, it excels in REM sleep detection, achieving a precision of 91.1%, a sensitivity of 84.7%, and an F1-Score of 87.5% in mice. These results underscore the advantages of using multimodal data and advanced neural architectures in sleep research. The main contributions of this work are summarized as follows:

(1) We propose a novel rodent automatic sleep classifier, MLS-Net, that demonstrates superior performance compared with other classifiers, achieving an overall accuracy rate of 90.4%.

(2) Our proposed model is based on a multimodal paradigm that takes multimodal signals (EEG, EMG, and EMs) as input. The employment of EM signals enhances the detection of REM sleep, achieving a precision of 91.1%, sensitivity of 84.7%, and an F1-Score of 87.5% in mice.

(3) We will open-source this multimodal dataset to foster broader research in sleep studies.

(4)   Our system achieves an inference time of 86.1 s for 24 h sleep recordings, which show strong potential for real-time applications.

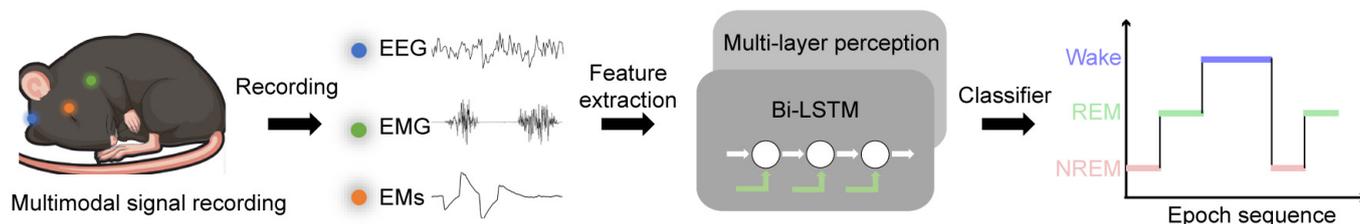A multimodal data based LSTM neural network for sleep stage classifications

Figure 1. The framework of a MLS-Net model using multimodal physiological signals. In MLS-Net model, the multimodal signals, including EEG, EMG, and EM signals, were collected from sleeping mice. The feature vectors of time–domain and frequency–domain were fed into bi-LSTM and then bi-LSTM using the feed-forward multi-layer perceptron (MLP) to compute the output probability of each sleep stage to identify the target of the signal.

2. Related Works

2.1. Different Approaches for ASSC in Human Dataset

Feature-based statistical machine learning approaches leverage pre-defined features as input. Classification accuracies vary significantly among the different classifiers reported in the literature, ranging from 70% to 94% [31]. Sanders et al. [32] proposed that a linear discriminant analysis (LDA) classifier achieved an average accuracy of 75% in ASSC using average power and cross-frequency coupling features extracted from EEG. Ra Boostani et al. [33] carried out a comparative review of several machine learning classification techniques used in ASSC and found that the RF classifier achieved a higher accuracy of 87.49% in healthy humans compared with KNN, SVM, and linear discriminant analysis (LDA). Similarly, Rahman et al. [34] reported that SVM (89.9%) and RF (90.2%) performed consistently better in terms of overall accuracy compared with RUSBoost on the SleepEDF dataset. Yoon et al. reported the encouraging performance of KNN with an accuracy of 98.65% using multi-modal signals, which has been adopted for sleep monitoring in a home-adapted device [35].

In recent years, deep neural networks, such as convolutional neural networks (CNNs) [36,37], recurrent neural networks (RNNs) [38], long short-term memory (LSTM) [39], deep belief networks (DBNs), and combinations of different network architectures [40,41], have been widely used in ASSC due to their capacity to handle large datasets effectively. For example (Table 1), Pei et al. [16] achieved an accuracy of 83.15% in ASSC using CNN and the gated recurrent unit (GRU) network on multichannel EEG data from the human sleep database. Integrating GNN with bidirectional GRU, Einizade et al. [42] reported an accuracy of 83.8% using multi-modal signals from the Sleep-EDF dataset. Awais et al. [43] designed a fusion model by integrating a DCNN with a SVM, achieving an accuracy of 93.8% in neonatal sleep stage classification using facial expression video data. The most impressive accuracy was reported by Mousavi et al. [44], who proposed an attention-based CNN-LSTM approach for sleep–wake detection using acceleration and heart rate variability and achieved 94.8% accuracy. However, the difficulty of collecting acceleration and heart rate data makes it unsuitable for further application.

Table 1. Summary of different sleep-stage algorithms in human.

Type	Classifier	Dataset	Channels	Accuracy	Ref.
Machine learning algorithm	Dendrogram multi-class SVM	15 subjects	EOG, EMG, EEG	88%	2015 [45]
	RF	Sleep–EDF dataset and CAP sleep dataset	EEG	87.49%	2017 [33]
	SVM RF RUSBoost	Sleep-EDF, Sleep-EDFX and ISRUC-Sleep databases	EOG	SVM: 89.9% RF: 90.2% RUSBoost:88.6%	2018 [34]
	KNN	10 subjects	PSG	98.65%	2022 [35]
	SVM; RF;	10 subjects	EEG, ECG, PPG, EOG	SVM: 70.05% RF: 70.75%	2023 [46]
	SleepBoost; CatBoost; GBoost; RF	20 recordings from Sleep-EDF dataset	EEG, EOG, EMG	86.30%	2024 [47]
Neural network	DNN	MIT-BIH database (18 subjects)	EEG, ECG	73.70%	2018 [48]
	RNN+CNN	Montreal Archive of Sleep Studies (200 subjects)	EEG, EOG, EMG.	87.10%	2019 [49]
	MLP	30 s segments from 19	EEG	82.53%	2020 [50]
	CNN	42 recordings from University Hospital of Leuven	EEG	64%	2020 [51]
	CNN+GRUs	SHHS (700 subjects)	EOG, EMG, EEG	83.15%	2020 [16]
	DCNN+SVM	19 Chinese neonates at the Children's Hospital of Fudan University	Vedio, facial expressions	93.80%	2021 [43]
	CNN + LSTM	Clinical dataset	HRV + acceleration	94.50%	2021 [44]
	GNN+bi-GRU	MASS; SleepEDF dataset	EOG, EMG, ECG, EEG	86.70% 83.80%	2023 [42]
	Traditional machine learning and deep learning techniques.	100 recordings from ISRUC and 197 recordings from Sleep-EDFx dataset	EEG, EMG, ECG	ISRUC dataset: 80.30% Sleep-EDF dataset: 81%	2024 [52]

Notation: RF: random forest; RUSBoost: random under-sampling boosting; SVM: support vector machine; CatBoost: categorical boost; RNN: recurrent neural network; CNN: convolutional neural network; MLP: multi-layer perceptron; LSTM: long short-term memory; DNN: deep neural network; DCNN: deep convolutional neural network; GRUs: gated recurrent units. ECG: electrocardiogram, EEG: electroencephalogram, EMG: electromyogram, EOG: electrooculogram, HRV: heart rate variability. PSG: polysomnogram.

2.2. Different Approaches for ASSC in Mice

The use of animal models is critical in preclinical research for circadian rhythm studies, disease modeling, and sleep monitor equipment development [23,53]. While ASSC has made significant progress in clinically utilizing multi-modal physiological signals, research in mice predominantly relies on single-modal signals such as EEG or EMG alone or a combination of EEG and EMG [23]. Various machine learning algorithms have been applied to rodent sleep classification, yielding promising results as shown in Table 2. While the success of these approaches has varied, the most effective methods have achieved approximately 92% accuracy in categorizing the three states of sleep–wake behavior [54]. For instance, Gross et al. [55] developed an open auto-scorer system based on complex Boolean logical decisions, achieving an accuracy of 80.24% using EEG and EMG signals in rats. Brankačk et al. [15] employed a linear discriminant analysis (LDA) classifier to predict vigilance states from 73 variables extracted from EEG data, attaining an accuracy of 89% in 10 mice. Rempe et al. [56] applied a hybrid approach combining principal component analysis (PCA) and a Naive Bayes classifier (NBC) for sleep classification using EEG and EMG signals. Although they achieved an overall accuracy of 90%, the accuracy for detecting REM sleep was only 70%. Svetnik et al. [13] compared the performance of the SVM algorithm with that of a deep neural network in mice, finding that the SVM achieved a comparable accuracy of 81%, only slightly lower than the 83% accuracy obtained with the deep neural network. In another study, Fraigne et al. [54] developed a novel ensemble learning approach, SleepEns, built on the Time Series Ensemble, which achieved 90% accuracy relative to expert scorers in mice. In contrast to the use of a single classifier, Gao et al. [12] proposed that employing multiple classifier systems is a more effective approach to improving the accuracy of automated sleep scoring.

Recent advancements have leveraged neural networks that utilize large amounts of training data and computational power to generate accurate predictions, eliminating the need for preliminary feature extraction. Exarchos et al. [57] applied a CNN-based

method, achieving a mean accuracy of approximately 85–90% for REM sleep detection. Tezuka et al. [30] developed UTSN-L, which processes single-channel EEG signals using a CNN combined with a LSTM network to incorporate historical information, reaching an overall accuracy of 90% in the mouse dataset. Yamabe et al. [58] generated the MC-SleepNet model using the same stratagem as UTSN-L and obtained 96.6% accuracy on a large-scale dataset. Pawan K. Jha et al. [59] created SlumberNet, which leverages the ResNet architecture for sleep scoring using EEG and EMG signals and excels with 97% overall accuracy. Although those state-of-the-art models achieved impressive accuracy in intra-lab validation, it also raises the concern of generalizing those models. Three publications [13,17,60] have described scoring algorithms validated across different datasets or species. To address the generalization challenges of ASSC in animals, Miladinović et al. [17] developed SPINDLE, a hybrid deep learning model incorporating a hidden Markov model (HMM) and CNN architecture. Empirical evidence indicates that SPINDLE effectively classifies sleep stages with an accuracy range of 93–99% while exhibiting robustness across diverse laboratory datasets. Similarly, Svetnik et al. [13] employed a CNN-based approach for sleep–wake scoring in non-human primates and other animals, achieving test accuracies of 75% for macaques, 83% for mice, 78% for rats, and 73% for dogs. Furthermore, Kam et al. [60] proposed the WSN, which utilizes wavelet-transformed images of mouse EEG/EMG signals. Their method achieved an accuracy of 86%, validated through a leave-one-subject-out (LOSO) approach on their dataset, as well as validation with an external dataset.

Table 2. Summary of different sleep-stage algorithms in mice.

Type	Classifier	Dataset	Channels	Validation	Accuracy	Ref.
Machine learning algorithm	Auto-Scorer	6 rats	EEG, EMG	Intra-lab	80.24%	2009 [55]
	LDA	10 mice	EEG	Intra-lab	89%	2010 [15]
	PCA/NBC	12 mice	EEG, EMG	Intra-lab	90%	2015 [56]
	RF	Rhesus macaques, Mice, Rat, Dog	EEG, EMG and ACT	Across-species	Macaques: 66% Mice: 81% Rat: 77% Dogs: 64%	2020 [13]
	SleepEns	28 mice	EEG, EMG	Intra-lab	90%	2023 [54]
Neural network	HMM + CNN	14 mice and 8 rats (SPINDLE dataset)	EEG, EMG	Cross-Lab	93–99%	2019 [17]
	CNN + LSTM	Large-scale dataset (4200 recordings); Small dataset (14 recordings)	EEG, EMG	Intra-lab	96.60%	2019 [58]
	CNN	Rhesus macaques, Mice, Rat, Dog	EEG, EMG and ACT	Across-species	Macaques: 75% Mice: 83% Rat: 78% Dogs: 73%	2020 [13]
	CNN	7 mice	EEG, EMG	Intra-lab	85–90%	2020 [57]
	CNN-LSTM	216 recordings (UNST-L dataset)	EEG	Intra-lab	90%	2021 [30]
	CNN	20 mice from intra-lab; AccuSleep dataset	Image dataset from EEG, EMG	Intra and cross-lab	Intra-lab: 86% Cross-lab: >85%	2021 [60]
	Residual Networks	5 mice	EEG, EMG	Intra-lab	97%	2024 [59]
	Pre-defined features and LSTM-MLP	7 mice from intra-lab; AccuSleep dataset	EEG, EMG, EMs	Intra-lab cross-lab	Intra-lab: 90.4% Cross-lab: 89.9%	Our work

Notation: LDA: linear discriminant analysis; PCA: principal component analysis; NBC: naive bayes classifier; RF: random forest; residual networks: based on a two-dimensional convolutional layer (Conv2D); ACT: locomotor activity; HMM: hidden Markov model.

2.3. Comparisons of MLS-Net with Prior Studies

In general, the poor generalization and suitability of feature-based statistical machine learning approaches result in inconsistent performance in ASSC [31], even when the same classifier is applied to the same dataset. Moreover, these approaches analyze each epoch independently, thereby failing to capture the time series information and the inherent rules of sleep transitions. Deep neural networks can partially address these shortcomings due to their enormous power to extract hidden information [21]. However, the non-interpretability of the model prediction and the longer computational times hinder the application of these models in real-time hardware implementations [20].

Our proposed MLS-Net integrates the strengths of feature-based approaches and deep neural networks. It uses the deterministic features extracted from multimodal physiological signals as input, which are then fed into a bi-LSTM network. These deterministic features enhance the interpretability of the model's performance. The LSTM network addresses the limitations of traditional machine learning methods (analyze each epoch independently) by processing time-series data and considering sleep transition rules. MLS-Net achieves optimal performance on our dataset with significantly reduced training time compared with other models. Most importantly, the lightweight nature of MLS-Net makes it highly feasible for real-time application in the future.

3. Materials and Methods

3.1. Animal

Male young C57Bl/6 J mice were purchased from Shanghai Jiesjie Laboratory Animal Co., LTD, China, and housed in rooms at 25 ± 1 °C and 50–70% humidity, under a 12/12 hr light/dark cycle (light on at 7 a.m.), with free access to food and water. For polysomnographic recording, the mice were surgically implanted with EEG and EMG according to previously established procedures [61]. All experiments were conducted in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and were approved by the Animal Care and Use Committee of Shanghai Medical College of Fudan University.

3.2. Multimodal Dataset Design and Implementation

For multimodal physiological signal recording, 7 healthy mice were recorded, and each mouse was recorded for 3 days according to the following procedure. A polysomnographic recording device was designed as previously described [61]. For EEG recording, two stainless steel screws were inserted on the top of left and right skulls at anteroposterior (AP) +1.5 mm, mediolateral (ML) +1.5 mm, and AP −3.5 mm, ML 3 mm, respectively. For EMG recording, two EMG electrodes were inserted into the neck musculature of the mice. Since freely moving mice close their eyes while sleeping, we cannot use cameras [62] or electrode eye coils to record Ems [63]. In our previous work, we innovatively proposed a novel eye movement tracking technique [61]. In brief, a miniature, strong magnetic rod was implanted in the conjunctiva of mice, and a magnetic displacement sensor was integrated into the polysomnographic device. The EMs of mice drive the sensor to cut through the magnetic field lines generated by the magnetic rod. The resulting change in the magnetic signal indicates the EMs of mice during sleep. Using this technique, we successfully achieved the simultaneous collection of EEG, EMG, and EM signals during mouse sleep.

To annotate the sleep–wake states of each epoch, raw EEG and EMG traces were visually inspected offline and scored into three vigilance states: wakefulness, NREM sleep, and REM sleep using SleepSign software (Version 2.0, Kissei Comtec, Nagano, Japan), followed by manual calibration by trained experts. The sleep states were segmented into non-overlapping consecutive 4 s epochs. Specifically, wakefulness epoch was characterized by increased EMG activity and low EEG amplitude. NREM sleep epoch was identified by delta-dominant EEG (0.5–4 Hz) and low EMG power. REM sleep epoch was defined based on theta-dominant EEG (6–9 Hz) and low EMG power.

3.3. MLS-Net Model

Our proposed MLS-Net algorithm consists of input preprocessing, feature extraction, bi-LSTM, multi-layer perception, and scoring output, as shown in Figure 2. Before implementing ASSC, a notch filter was applied to eliminate 50 Hz noise in EEG, EMG, and EMs. The EMG signal was subsequently processed using a Butterworth bandpass filter with a frequency range of 30–70 Hz to effectively eliminate significant electrical noise. Subsequently, extreme outliers in EEG and EMs caused by motion artifacts and EMG contamination were removed as follows. We defined the extreme lower tail percentile at 0.01 th and extreme

higher tail percentile at 99.99 th in EEG and EM signals. The data exceeding the thresholds were set to the threshold values and then normalized using the following formula. Subsequently, the time and frequency domain features were extracted from outlier removal and normalized multimodal physiological signals. The feature vectors were fed into a bi-LSTM layer, which was a sequence-processing neural network that incorporated past states (i.e., epochs) into the classification of present states and captured temporal dependencies. The bi-LSTM layer comprises 16 units in each direction, totaling 32 units. The outputs from the bi-LSTM layer are then fed into a multi-layer perceptron (MLP), followed by a softmax layer, which generates class probabilities.

$$Signal_{Normalized} = \frac{Signal - \mu_{Signal}}{\sigma_{Signal}}$$
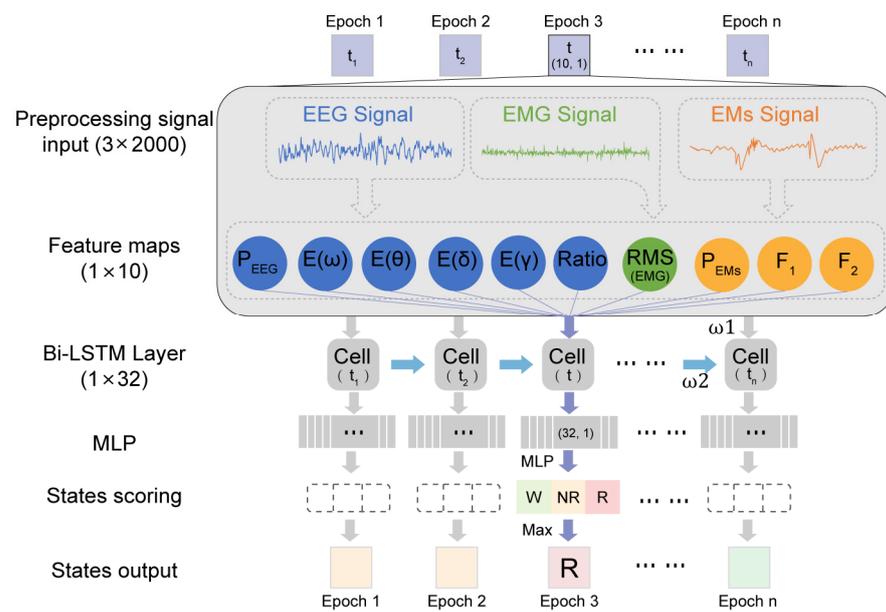

Figure 2. Illustration of MLS-Net model.

Signal refers to multi-modal signals. μ_{Signal} is the mean of the raw signal, and the δ_{Signal} is the standard deviation of the raw signal. This formula is general for EEG, EMG, and EMs by changing the subscript "*Signal*' to the respective signal type.

The neural network models were trained using an Intel Core i7-11800H processor, 16 GB of memory (8 GB × 2), and an NVIDIA GeForce RTX 3060 GPU.

3.4. Expert-Knowledge-Based Feature Extraction

For pre-defined features, all methods tested in this paper start with the identical feature extraction step, where a wide range of features were computed from the pre-processed EEG, EOG, and EMG signals (linear time and frequency domains and non-linear time–domain features). Frequency domain features of EEG signals were performed using Fast Fourier Transform (FFT) algorithms. The energy of EMG, which carries information about muscle activity, was also calculated. The frequency of EMs was determined according to our previous research [61]. A total of 10 features were derived for feature-based analysis, including the power of regular waves (0.5–4 Hz, 6–9 Hz, 8–12 Hz, and 52–70 Hz), the root mean-squared of EMG, the mean amplitude of EEG and EMs, and theta/delta EEG power ratio.

(1)   Amplitude and Power Calculation for EEG Signals:

$$Power_{signal}(\omega) = \sum_{t=0}^{T} |F(\omega, t)|^2 \text{ for } \omega \epsilon [0.5\text{Hz}, 100\text{Hz}]$$

(1)

where *F(w, t)* is the FFT of the signal at time t and frequency ω, and *T* is the total duration. This formula is general for EEG, EMG, and EMs by changing the subscript "signal" to the respective signal type.

(2)   Spectral Power Features for EEG:

$$P_{EEG}(band) = \frac{\sum_{t=start}^{end} \left[\sum_{f=f_{low}}^{f_{high}} |F(f, t)|^2 \right]}{end - start + 1}$$

(2)

where band specifies the frequency band (e.g., $P_{EEG}(\delta) \epsilon$ [0.45 Hz, 4 Hz], E(θ) ϵ [6 Hz– 9 Hz], upper E(θ) [7 Hz, 8.5 Hz], E(α) ϵ [8 Hz, 12 Hz], $P_{EEG}(\omega) \epsilon$ [0 Hz, 30 Hz], upper E(γ) ϵ [52 Hz, 70 Hz]). The upper E(θ) and upper E(γ) have been proven to be effective for discriminating between sleep and wake [15]. f_{low} and f_{high} are the boundaries of the frequency band, and start and end define the time window.

(3)   Theta/Delta Ratio and Sleep State EM Features:

$$Ratio_{\theta/\delta} = \frac{P_{EEG}(theta)}{P_{EEG}(delta)}, \quad F_{EMs}(state) = f_{EMs}(A_{state}) \text{ for } state \epsilon (wake, REM)$$

(3)

where $P_{EEG}(theta)$ and $P_{EEG}(delta)$ are power computations in the theta (6–9 Hz) and delta (0.45–4 Hz) bands, respectively, and A_{state} denotes the threshold values (mean $\pm$ 1.96 × std.) of EMs during different periods (wakefulness or REM). $F_{EMs}(state)$ represent the frequency of EMs exceeding A_{wake} and A_{REM} thresholds in the current epoch.

(4)   The root mean squared (RMS) of EMG:

$$RMS = \sqrt[2]{\frac{1}{n} \sum_{i=1}^{N} x_i^2}$$

N represents the total number of data points in the EEG signal. x_i denotes the *i*-th data point of the signal.

3.5. Evaluation

In order to evaluate the training and judgment accuracy of different models, all algorithms should be developed using the same channels, trained on the same pre-clinical dataset, and validated with the same procedure. In our algorithm, the effectiveness of different algorithms was validated by leave-one-out cross-validation (LOOCV). Out of all the "k" recordings derived from 7 mice, a single record was retained as the validation data for testing the model, and the remaining k-1 records were used as training sample. The LOOCV process was repeated k times and yielded k results, with each of the k records used exactly once. The average of k results was deemed to be a single estimation for each algorithm. Using these strategies, all records were used for both training and validation, so that the evaluation bias due to single training could be effectively avoided without compromising the stability and reproducibility of the algorithms. Five different evaluation measures are used to evaluate the performance of the proposed method, including the confusion matrix, accuracy, precision, sensitivity (recall), and F1-Score, as shown in Figure 3 and Table 4 [64].

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}\% \quad Precision = \frac{TP}{TP + FP}\%$$

(4)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \% \text{F1} - \text{Score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$
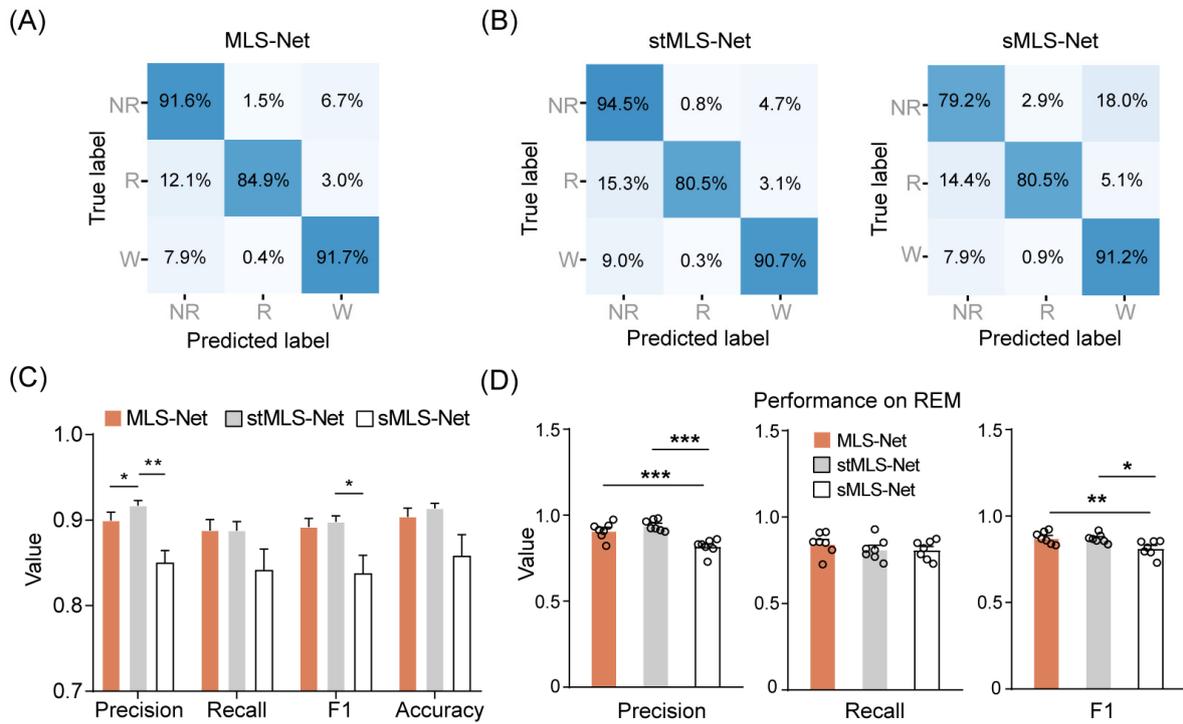

Figure 3. Performance of the MLS-Net for sleep-wake classification. **(A)** Confusion matrix of MLS-Net evaluated on the test set. **(B)** Confusion matrix of stMLS-Net and sMLS-Net evaluated on the test set. Spatio–temporal MLS-Net model (stMLS-Net) and spatial MLS-Net model (sMLS-Net) were variants based on MLS-Net model. Each element in the matrices indicates the percentage of correct matching between experts (True label) and each model (Predicted label). The abbreviations used are as follows: wake (W), eye movement sleep (R), non-eye movement sleep (NR). **(C)** Overall performance of MLS-Net, stMLS-Net and sMLS-Net models. **(D)** Precision, sensitivity (recall), and F1-Score of MLS-Net, stMLS-Net and sMLS-Net models for REM stage classification. Empty circle in the histogram indicates individual result in cross-validation. One-way ANOVA followed Dunnett's post hoc analysis. Data were presented as mean ± SEM. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

In this context, TN, FN, FP, and TP refer to true negatives, false negatives, false positives, and true positives, respectively. TN represents the number of sleep stages that were incorrectly classified as corresponding to the labeled sleep stages. FN denotes the number of sleep stages that were inaccurately identified as other stages when they should have been classified as the correct stage. FP refers to the number of sleep stages that were erroneously classified as labeled stages when they were not. TP indicates the number of sleep stages that were correctly classified in accordance with their labeled stage.

The coefficient of variation (CV) within groups is calculated using the following formula:

$$CV = \frac{\sigma}{\mu} \times 100\%$$

In this formula, σ is the standard deviation of each accuracy for all-fold cross-validation, μ is the mean of the performance accuracy.

3.6. The Construction of Rebalance Dataset Using Data Augmentation

To address the class imbalance issue, we performed data augmentation by repeating the minority class samples as suggested in previous studies [65]. The raw epoch sequences comprise varying durations for wake, NREM, and REM epochs. Initially, we calculated

the mean length of epochs within each continuous class. Subsequently, we trimmed all epochs exceeding this mean length to the calculated value. We then increased the number of samples for the minority classes, expanding them to match the mean duration of the wake, NREM, and REM classes in the input sequence. This approach enabled us to construct an augmented dataset, referred to as the balanced dataset, ensuring that all sleep stages are equally represented in the training set. The augmentation process is illustrated in Figure 6A.

4. Result

4.1. Dataset

From the animal experiments, a small-scale sleep recording dataset was compiled. A total of seven recordings from seven healthy wild-type mice were conducted. Each recording lasted three days, with data from the last 24 h incorporated into the dataset. As shown in Table 3, the multimodal dataset comprises 151,200 epochs, including 82,449 wake epochs (54.5%), 56,286 NREM epochs (37.2%), and 12,465 REM epochs (8.2%). The proportions of sleep architecture are consistent with previous reports [66]. Each epoch includes one channel of EEG, EMG, and EM signals, simultaneously collected with a duration of 4 s at a 500 Hz sample rate. Prior to feature extraction, preprocessing was conducted, which included 50 Hz noise filtering and the removal of outliers in artifact-contaminated epochs, following the predefined rules outlined in the methods section. For inner-cross validation, the dataset was split into training and validation sets, with 85.7% of the sleep samples used for training and the remaining 14.3% for validation.

Table 3. The distribution of sleep stages in the dataset.

Stage	Total Epochs	Percentage (%)	Training Set (%)	Test Set (%)
Wake	82,449	54.5		
REM	12,465	8.2	85.7 (6/7)	14.3 (1/7)
NREM	56,286	37.2		

4.2. The Performance of MLS-Net and Its Variants

We designed the MLS-Net to process multi-modal signals from our constructed dataset. To minimize the variance of predictions across different mouse recordings, we employed a cross-validation approach using seven separate MLS-Net trained on a Leave-One-Out Cross-Validation (LOOCV) method. This approach ensured that six animal recordings were used for training, while the remaining one, which was never seen by the model during training, was used for validation. Figure 3A shows the average confusion matrix comparing the true and predicted sleep labels for the input test set after the 120 th iteration of training on the MLS-Net model. Our proposed MLS-Net model achieves favorable diagonal coefficients in its confusion matrix, with an accuracy of 91.6% for NREM, 84.9% for REM, and 91.7% for wake, respectively. and yielded an overall decoding precision of 89.9%, a mean sensitivity (recall) of 88.8%, an F1-Score of 89.2%, and an overall accuracy of 90.4% (Table 4, Figure 3C).

Table 4. Performance comparison of the different models.

Algorithms	Dataset	EM-Exclusion	Balance	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Time (s)
MLS-Net	Our dataset	No	No	89.9	88.8	89.2	90.4	86.1
		No	Yes	74.2	79.5	81.5	81.2	
		Yes	No	80.8	76.8	78.4	80.6	
	AccuSleep Dataset	--	No	90.0	85.0	87.1	89.9	
stMLS-Net	Our dataset	No	No	91.6	88.9	90.0	91.4	3073.3
			Yes	78.4	72.7	74.9	75.7	
sMLS-Net	Our dataset	No	No	84.3	83.6	83.9	85.8	479.8
			Yes	74.1	75.4	74.7	76.4	
***MC-SleepNet**	Our dataset	No	No	88.5	84.8	86.4	89.0	608.9
			Yes	78.6	70.7	73.5	74.6	
***SlumberNet**	Our dataset	No	No	88.0	86.5	86.0	86.5	3178.4
			Yes	80.5	79.8	79.3	79.6	

Table 4. *Cont.*

Algorithms	Dataset	EM-Exclusion	Balance	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Time (s)
RF	Our dataset	No	No	84.5	78.9	78.3	**83.3**	2.9
			Yes	78.1	73.4	75.3	68.2	
LR	Our dataset	No	No	83.9	80.8	82.2	**73.1**	10.7
			Yes	76.3	72.0	73.8	67.7	
XGBoost	Our dataset	No	No	73.4	59.5	48.0	**83.6**	22.1
			Yes	76.5	73.4	74.4	67.3	

Notation: AccuSleep dataset was an external dataset (https://doi.org/10.17605/OSF.IO/PY5EB accessed on 22 July 2024). EMs-exclusion means whether the EMs signals were removed during the training sets. AccuSleep dataset was devoid of EM signals, so this section was left unfilled. The balance means whether the data augmentation were conducted. All performance metrics (precision, sensitivity (recall), F1-Score, and accuracy) are averaged values across LOOCV validation. All models were tested on both balanced and imbalanced datasets. The accuracy values for the imbalanced dataset (raw dataset) are highlighted in bold. Time of each model indicates the compute time after 120 iterations, with each classifier trained on six recordings.

The issue of network structure is crucial for understanding the trade-offs between model complexity and its practical application, especially in real-time systems such as sleep state analysis. To explore the appropriate structure of the model, we further generated two variants on the basis of MLS-Net. In the spatial MLS-Net model (sMLS-Net), the pre-defined features were removed and replaced by a two-layer CNN for spatial feature extraction from EEG, EMG, and EM signals, followed by an MLP layer and a softmax layer. The spatio–temporal MLS-Net model (stMLS-Net) reserved the pre-defined features, which were concatenated with spatial features extracted from CNN as input. The series information was also captured by a bi-LSTM neural network. These are then processed through a MLP, and finally, a softmax layer outputs three targets. These two classifiers were trained and tested using the identical data partitions (six training sets versus one testing set) of our dataset.

Compared with MLS-Net, the stMLS-Net model shows a competitive performance on wake (90.7%) and NREM (94.5%), but the sMLS-Net model had lower accuracy both in NREM (79.2%) and REM (80.5%), as shown in Figure 3B. The overall performance of MLS-Net and stMLS-Net is comparable in terms of precision, sensitivity (recall), and F1-Score, with both models outperforming the sMLS-Net model in these metrics (Figure 3C). This result demonstrated the importance of the LSTM layer in MLS-Net, as the sMLS-Net, which lacks the LSTM layer, shows poor performance, especially in REM sleep classification (Figure 3D).

4.3. Performance Comparison of Various Classifiers

To evaluate the performance of MLS-Net, we conducted a comparative analysis against five other classifiers. Among these, two represent state-of-the-art models based on neural networks, specifically MC-SleepNet [58] and SlumberNet [59], while the remaining three are machine learning models: random forest (RF), linear regression (LR), and XGBoost. To ensure a fair comparison, all models were trained and tested on the same dataset, following identical preprocessing steps. Notably, MC-SleepNet and SlumberNet were well-deigned to process EEG and EMG signals, whereas the other three models were meticulously adjusted to accommodate EM signals as input. The predictive performance of these models is summarized in Figure 4A and Table 4.
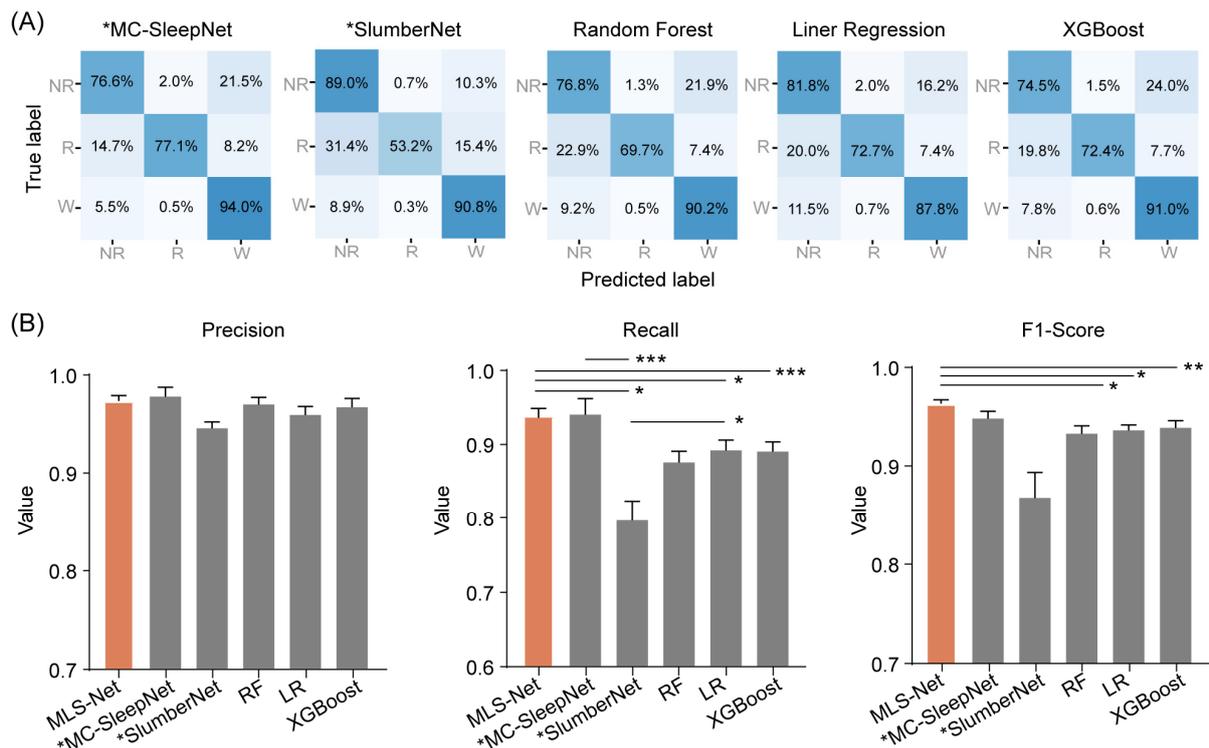
Figure 4. The performance of different classifiers on REM classification. (**A**) Shows the confusion matrix of MC-SleepNet, SlumberNet, random forest (RF), linear regression (LR), and eXtreme Gradient Boosting (XGBoost). The state-of-the-art models MC-SleepNet and SlumberNet were denoted with an asterisk (*MC-SleepNet, *SlumberNet). Each element in the matrices indicates the percentage of correct matching between experts (True label) and each model (Predicted label). The abbreviations used are as follows: wake (W), eye movement sleep (R), non-eye movement sleep (NR). (**B**) shows the performance comparison of precision, sensitivity (recall), and F1-Score for REM classification. The orange bar indicates the performance of MLS-Net model and the gray bars indicate the performance of other models. One-way ANOVA followed Dunnett's post hoc analysis. Data were presented as mean $\pm$ SEM. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

The confusion matrix reveals that both algorithms exhibit excellent performance in classifying the wake but perform poorly in distinguishing NREM and REM stages, especially in the case of SlumberNet and feature-based machine learning algorithms (Figure 3D). From the confusion matrix, we can see that the primary source of misclassification for NREM stages is their erroneous attribution to the wake stage, while misclassification of REM stages mainly arises from their incorrect attribution to NREM stages, as shown in the hypnograms in Figure 5A,B. In terms of overall accuracy, the MLS-Net model achieved the highest accuracy (89.4%), followed closely by the MC-SleepNet (89.0%). The SlumberNet, RF, LR, and XGBoost models achieved accuracies of 86.5%, 83.3%, 73.1%, and 83.6%, respectively (Table 4). In addition to performance, we also compared the training time consumed by each classifier. We found that the MLS-Net model had a significantly faster training time of 86.1 s per 120 iterations over the same training set than its two variants (stMLS-Net, sMLS-Net) and MC-SleepNet (608.9 s). In contrast, SlumberNet requires a longer training duration, up to 3718.4 s for each LOOCV.
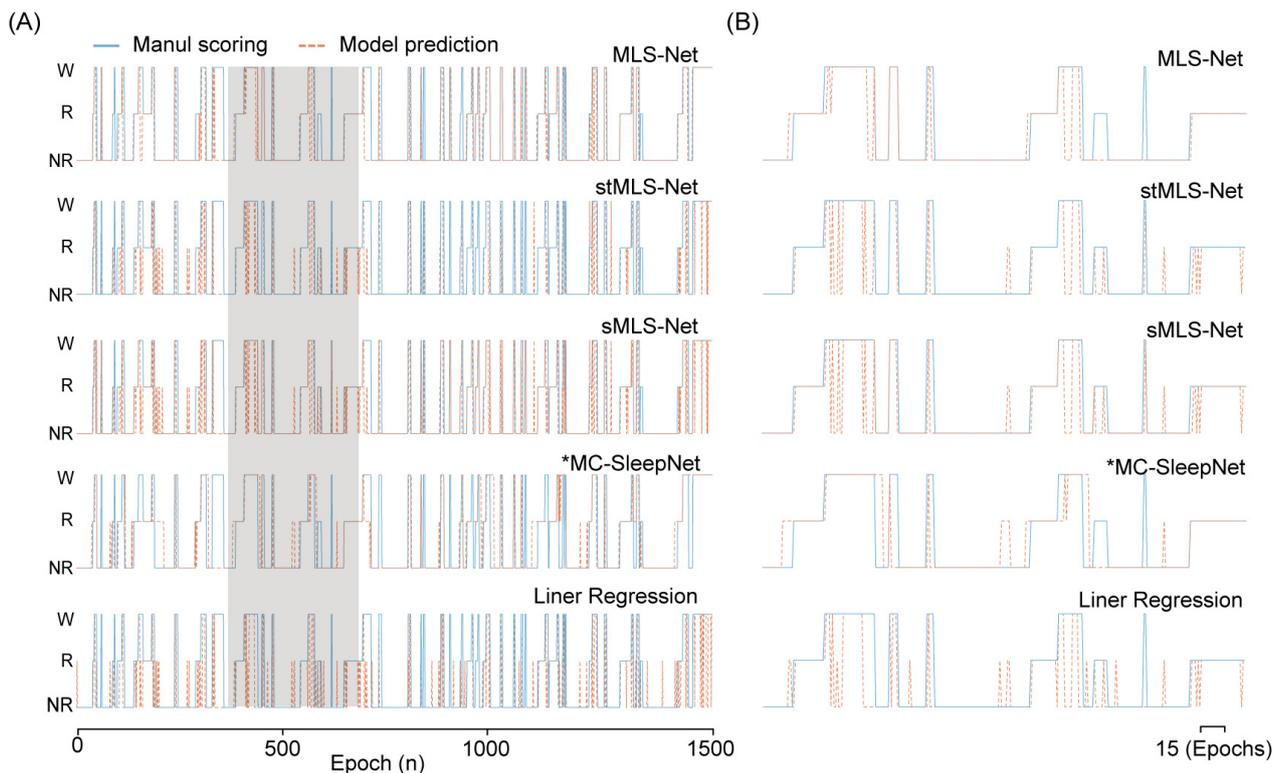
Figure 5. The output hypnogram predicted by different models compared with manual scoring. (**A**) shows example hypnograms of one subject. The solid blue line and yellow dashed line denote the hypnograms depicted by the proposed model and an expert, respectively. (**B**) was the enlargement of hypnogram in the shielding area (**A**).

Further inspection shows that MLS-Net achieved a precision of 92.1%, a sensitivity (recall) of 83.4%, and an F1-Score of 87.3% for REM sleep (Figure 4B). The sensitivity of MLS-Net in REM classification outperforms SlumberNet and traditional machine learning algorithms (Figure 4B). This improvement can be attributed to the utilization of EM signals, which enhance the discrimination between REM and NREM sleep stages.

4.4. The Performance of Different Classifiers on Balanced Dataset

The sleep stages identified by the expert occur with widely differing frequencies [65] (e.g., wake: 54%, REM: 8.2%, NREM: 37.2% in our dataset). Such differences in the frequencies of classes (class imbalance) can potentially hinder the performance of classifiers and limit further advancements in sleep staging algorithms [65,67]. To address this issue, we employed a data augmentation technique by duplicating minority class samples in the sleep epoch sequence (Figure 6A). In the balanced dataset, all samples had an equal opportunity to compensate for the minority classes. Next, we trained and tested the performance on both rebalance datasets. After 120 iterations on the training dataset, the overall accuracy of the MLS-Net, sMLS-Net, and SlumberNet models on the balanced dataset decreased by 9.2% (from 90.4% to 81.2%), 9.48% (from 85.8% to 76.4%), and 6.9% (from 86.5% to 79.6%), respectively (Figure 4B). Although stMLS-Net and MC-SleepNet exhibited comparable performance to the MLS-Net model on the standard dataset (Tabel 4), their accuracy on the balanced dataset dropped even more significantly, with declines of 15.7% (from 91.4% to 75.7%) and 14.4% (from 89.0% to 74.6%), respectively (Figure 4B). These results suggest that both the spatio–temporal MLS-Net (stMLS-Net) and MC-SleepNet networks are more susceptible to imbalanced tasks in this work. Moreover, MLS-Net consistently outperformed the other algorithms with relatively lower variance for each model of the LOOCV cross-validation (Figure 6D).
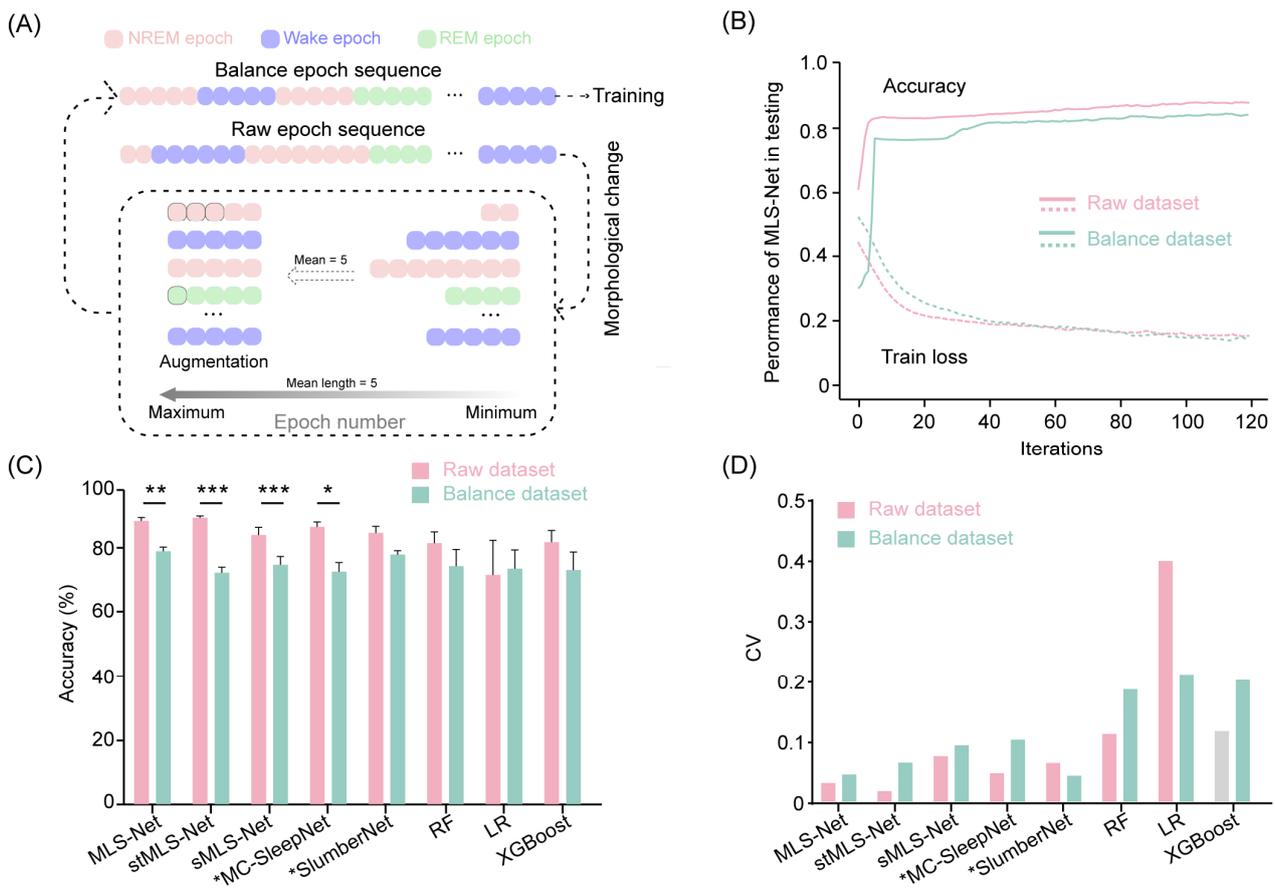
Figure 6. Accuracy comparison cross-models in raw dataset and balanced dataset. (**A**) The generic workflow of data augmentation. (**B**) Accuracy and loss curves of the training and testing on raw dataset and balanced dataset. (**C**) The scoring accuracy of different models for all folds in cross-validation. Two-sided paired Student's t-test between raw dataset and balanced dataset for each model. Data were presented as mean ± SEM. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. (**D**) The coefficient of variation (CV) of accuracy for all folds in cross-validation.

4.5. The Influence of EM Signals on MLS-Net Model

Overall, our proposed MLS-Net model demonstrates performance that is equal to or superior to that of state-of-the-art models such as MC-SleepNet, SlumberNet, and other machine learning algorithms on both standard and imbalanced datasets. One possible reason for this enhanced performance is the use of EM signals. To evaluate the contribution of EM signals to the performance of MLS-Net, we performed EM-exclusion testing, where features extracted from EMs were replaced with a null mask. The class-specific accuracy was evaluated using a mean confusion matrix and overall accuracy metrics, as illustrated in Figure 7A,B.

The exclusion of EM signals resulted in agreement rates of 77.9% and 72.2% for REM and NREM sleep, respectively, between predicted classifications and true labels. An increased incidence of erroneously categorizing REM and NREM epochs as wake epochs was observed, as depicted in Figure 7A. Consequently, the overall accuracy of sleep state scoring significantly decreased from 90.4% to 81.2% when EM signals were excluded. (Figure 7B). A closer investigation has shown that precision, sensitivity (recall), and F1-Score all decreased for individual stages after EMs-exclusion (Figure 7C). Specifically, the F1-Score for wake significantly declined from 92.7% to 84.6% after EMs-exclusion, as illustrated in Figure 7C. These results indicate that incorporating EM signals can significantly enhance the performance of the MLS-Net by providing additional information during training.
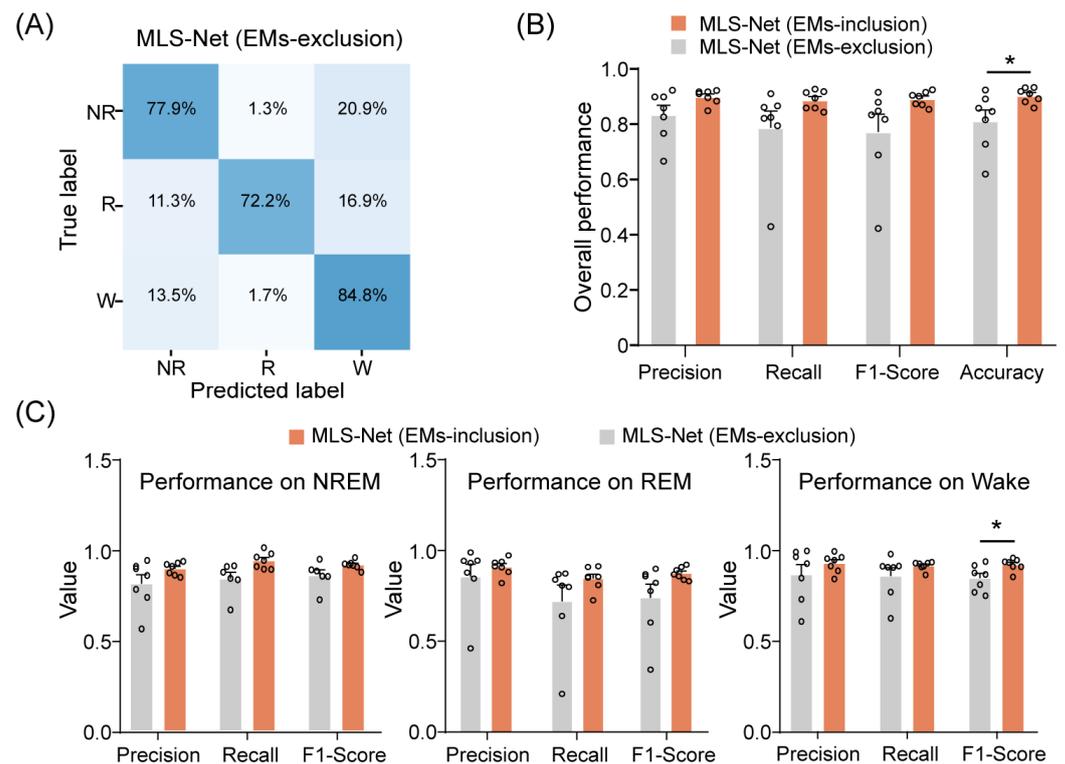
Figure 7. The performance of MLS-Net models after EMs-exclusion. (**A**) The confusion matrix of MLS-Net performance on the EMs-exclusion dataset. (**B**) Overall accuracy of MLS-Net performance across all folds of cross-validation. (**C**) Precision, sensitivity (recall), and F1-Score for individual stage classification following EMs-exclusion. Empty circle in the histogram indicates individual result in cross-validation. Two-sided Student's *t*-test. Data were presented as mean $\pm$ SEM. * $p < 0.05$.

4.6. External Validation of MLS-Net Model

To evaluate the generalizability of MLS-Net, we conducted an external validation using the AccuSleep mouse EEG/EMG sleep dataset (https://doi.org/10.17605/OSF.IO/PY5EB, accessed on 22 July 2024). This dataset includes sleep recordings from five healthy mice (mouse01–mouse05), each consisting of two complete 24 h polysomnography (PSG) records. The PSG data comprise one EEG and one EMG signal, sampled at 512 Hz, with sleep states annotated by experts every 2.5 s. We utilized a continuous 24 h segment from each mouse to validate the performance of the MLS-Net model. To ensure a fair comparison, all data were processed within the Pytorch framework, and the parameters of the baseline model were maintained as previously specified.

Since the AccuSleep dataset does not include EM signals, we left the corresponding EM parameters (Feature 8, 9, and 10) blank, as shown in Figure 8A. Apart from the features extracted from EM signals, the distributions of the multimodal signal features extracted from the external dataset were similar to those extracted from our dataset, as shown in Figure 8A. From a sleep perspective, we further compared the proportions of NREM, REM, and wake epochs between the external dataset and our dataset. The external dataset had proportions of 42.6% for NREM, 6.4% for REM, and 50.9% for wake, which were analogous to the proportions in our dataset (average 37.2% for NREM, 8.2% for REM, and 54.5% for wake) in Figure 8B.
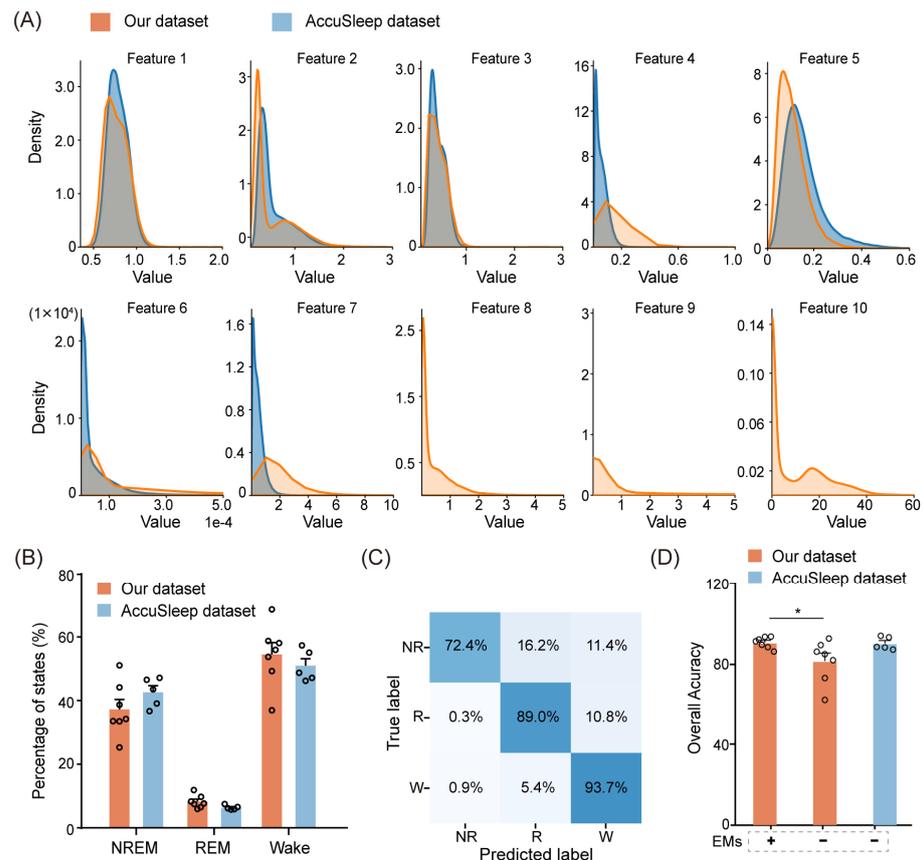
Figure 8. Performance of MLS-Net on an external dataset. (**A**) The distribution of pre-defined features extracted from each epoch in our dataset (yellow) and the external dataset (blue). (**B**) A comparison of sleep architecture between our dataset and the external dataset. (**C**) The confusion matrix demonstrating the performance of MLS-Net on the external dataset. (**D**) Overall accuracy of MLS-Net performance on the external dataset. The symbols '+' and '−' indicate testing on the EMs included and EMs excluded datasets, respectively. Empty circle in the histogram indicates individual result in cross-validation. One-way ANOVA followed Dunnett's post hoc analysis. Data were presented as mean $\pm$ SEM. * $p < 0.05$.

The performance of MLS-Net in the AccuSleep dataset was shown in the confusion matrix and overall accuracy against consensus manual scores. As shown in Figure 8C, the agreement between MLS-Net output and human score on the AccuSleep dataset has reached 93.7% and 89.0% for wake and REM stage recognition, respectively. The mean accuracy on the AccuSleep dataset was 89.9%, which is slightly lower than the accuracy of 90.4% achieved on our dataset (Figure 8D). The accuracy of 89.9% in the AccuSleep dataset, exceeding 80.6% achieved in the EMs-exclusion test on our dataset, can be attributed to two main factors. Firstly, the high quality of EEG and EMG signals in the external dataset enabled the model to capture sufficient information for parameter optimization. Secondly, the performance of a neural network is correlated with the amount of training data [28]. Therefore, the shorter duration of sleep state epochs (2.5 s/epoch vs. 4 s/epoch) in the external dataset allowed for a greater number of epochs to be extracted from the same amount of sleep data for training. In summary, MLS-Net successfully withstood external data validation, thereby further enhancing the model's generalizability.

5. Discussion

5.1. Hybrid of Feature-Based and Neural Network Methodology

Over the past 30 years, automatic sleep scoring has seen significant advancements. Compared with feature-based methodologies, deep neural networks demonstrate remark-

able efficacy in extracting latent information from raw physiological signals. However, its "black box" nature makes the information difficult to characterize and translate into the clear, interpretable features used in feature-based approaches [21]. The lack of interpretability in results and prolonged computational times are the main drawbacks that could limit acceptance by end users [68]. In contrast, feature-based approaches provide clear insights into the sleep characteristics considered during scoring. Traditional feature-based approaches also present challenges, as they analyze each input segment independently and do not fully exploit the temporal sequence information inherent in sleep–wake transitions. The bi-LSTM neural network, with its forget gate structure, widens the usage of past information and is better suited for time series data [64]. In this work, we leverage the strengths of both feature-based approaches and bi-LSTM neural networks to develop the MLS-Net model, which utilizes deterministic expert knowledge to enhance the performance of neural networks. The performance of MLS-Net is validated by LOOCV in mice. This internal cross-validation ensures that data from the test set does not appear in the training set, thereby enhancing the model's generalizability to unfamiliar datasets.

The overall decoding precision, sensitivity, F1-Score, and accuracy of MLS-Net at the testing set are 89.9%, 88.8%, 89.2%, and 90.4%, respectively. After training with a cross-validated approach, we performed a head-to-head comparison and found that the MLS-Net model demonstrates superior performance over the state-of-the-art MC-SleepNet [58], SlumberNet [59], and traditional maching learning models with higher accuracy and the lowest CV when using the same channels (Figure 6). Due to the relatively small proportion (5–10%) of REM epochs within the entire sleep–wake cycle, accurately identifying REM sleep poses a significant challenge. Our findings demonstrate that MLS-Net achieved an 84.9% agreement with expert scorers on REM epoch classification, while MC-SleepNet, SlumberNet, random forest (RF), linear regression (LR), and XGBoost achieved agreement rates of 77.1%, 53.2%, 69.7%, 72.7%, and 72.4%, respectively. Meanwhile, the promising and stable performance of MLS-Net was validated using an external dataset, where it maintained an accuracy of over 89% even without the use of EM signals. The notable performance in REM sleep classification can be attributed to the incorporation of deterministic features extracted based on expert knowledge from multi-modal signals and the temporal information captured by the bi-LSTM layers.

5.2. EM Signals during REM Sleep Contributes to Sleep Classification

EMs in sleep were originally characterized as falling into the non-rapid eye movement sleep (NREM) and rapid eye movement sleep (REM) categories [69]. REM sleep is characterized by frequent EMs, whereas NREM sleep is marked by the relative absence of such movements. This phenomenon is highly conserved across different species, from humans to rodents [61,70]. Previous studies reported that the average frequency of EMs in mice was about 0.90, 0.25, and 0.05 Hz during wake, REM, and NREM sleep, respectively [71]. Thus, EMs could be a potential indicator to distinguish between different stages in mice. One non-invasive method for detecting EMs is through EOG recording, which has been shown to effectively enhance the performance of sleep classification algorithms in clinical settings [34]. However, the intricate technical challenges have impeded its implementation in small animals [72]. Our previous work [61] enabled the simultaneous collection of EEG, EMG, and EM signals from freely moving mice. In this study, we deigned MLS-Net based on three modalities of signal fusions (EEG/EMG/EMs). By performing EMs-exclusion testing, we found that EMs significantly improved the performance of MSL-Net in discriminating sleep status with high accuracy, which was paralleled with counterpart models.

5.3. Limitations and Future Perspectives

Although the MLS-Net model had substantial agreement with sleep experts, there are important limitations to consider. While we tested the model with an external mouse sleep dataset, the robustness and generalizability of MLS-Net have not yet been validated for human use. Human PSG records involve a more complex classification into multiple sleep

stages compared with rodents, which may pose greater challenges [23]. Much research has been done on improving human EEG classification as well [12]. MLS-Net capitalizes on the strengths of neural networks in capturing temporal sequence information while incorporating predefined features extracted from multimodal signals such as EEG, EMG, and EMs from mice. Due to the significant differences in sleep architecture and feature distributions between mice and humans, our findings cannot be directly extended to human studies. However, given that multi-modal electrophysiological data can be collected not only in mice but also in humans [73], adapting the MLS-Net model for human datasets could involve several potential modifications. These may include adjusting the frequency ranges used in feature extraction to better align with human EEG frequencies or fine-tuning the memory capacity of LSTM layers to more effectively capture the long-term dependencies characteristic of human sleep patterns. Secondly, the outstanding performance of the MLS-Net model was validated on a small-scale dataset, but its performance on large-scale datasets remains unknown. In theory, a larger amount of training data would lead to a better training effect for the deep learning method [28]. Additionally, our study was limited in exploring potential variations in accuracy due to factors such as age, behavior, gender, or genetic variability among mice.

Given the inherent advantages and limitations of various preprocessing methodologies, an optimal ASSC system may be derived from a judicious amalgamation of different techniques [10]. Future research in deep learning is expected to prioritize the development of simpler models for ASSC that maintain high accuracy while optimizing portability and efficiency. Efforts should also focus on extracting channel-specific information and identifying micro-stages within NREM and REM sleep without compromising algorithmic stability. Moreover, enhancing model compatibility is crucial to facilitate its application in sleep classification for the pathological states of rodents, potentially aiding in the screening and identification of physiological markers for sleep-related mental illnesses in preclinical research.

6. Conclusions

This study designed the MLS-Net model that combines the strengths of neural networks and feature extraction for rodent ASSC using multimodal biological signals (e.g., EEG, EMG, EMs). By leveraging explicit feature inputs, the proposed MLS-Net model effectively mitigates the overfitting commonly associated with neural networks and addresses the low accuracy issues typically encountered in feature-based machine learning methods. The MLS-Net model has a runtime of only 86.1 s for 24 h sleep recordings, which is significantly faster than the several hundred to thousands of seconds required by other networks. The time-saving property without compromising its efficiency makes MLS-Net suitable for real-time sleep scoring applications in preclinical research. Moreover, the algorithm offers flexibility for users who may wish to customize it to better suit their specific needs. Potential modifications include using alternate features, adjusting artifact detection criteria during preprocessing, or incorporating additional signals to enhance performance.

Author Contributions: C.J., J.L. and J.Z. (Jiayi Zhang) conceived and designed the research; C.J., W.X. and B.Y. designed the machine learning algorithm; C.J. and J.Z. (Jiadong Zheng) conducted the experimental study; C.J., J.L. and J.Z. (Jiayi Zhang) wrote the paper. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset and code utilized in this study are openly accessible. Specifically, the MLS-Net model code referenced in this article is available on Kaggle at the following

link: https://www.kaggle.com/models/jiangchengyong/mls-net (accessed on 30 May 2024). Additionally, the multimodal physiological dataset constructed from mice is also uploaded on Kaggle and can be accessed here: https://www.kaggle.com/datasets/chenruilin245/sleepdata (accessed on 30 May 2024). Furthermore, the publicly available AccuSleep dataset, which was analyzed in this study, can be found at: https://doi.org/10.17605/OSF.IO/PY5EB (accessed on 30 May 2024).

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

References

1. Girardeau, G.; Lopes-Dos-Santos, V. Brain neural patterns and the memory function of sleep. *Science* **2021**, *374*, 560–564. [CrossRef]
2. Zoubek, L.; Charbonnier, S.; Lesecq, S.; Buguet, A.; Chapotot, F. Feature selection for sleep/wake stages classification using data driven methods. *Biomed. Signal Process. Control* **2007**, *2*, 171–179. [CrossRef]
3. Redmond, S.J.; Heneghan, C. Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 485–496. [CrossRef]
4. Riemann, D.; Nissen, C.; Palagini, L.; Otte, A.; Perlis, M.L.; Spiegelhalder, K. The neurobiology, investigation, and treatment of chronic insomnia. *Lancet Neurol.* **2015**, *14*, 547–558. [CrossRef]
5. Toth, L.A.; Bhargava, P. Animal models of sleep disorders. *Comp. Med.* **2013**, *63*, 91–104.
6. Bixler, E.O.; Kales, A.; Vela-Bueno, A.; Drozdiak, R.A.; Jacoby, J.A.; Manfredi, R.L. Narcolepsy/cataplexy. III: Nocturnal sleep and wakefulness patterns. *Int. J. Neurosci.* **1986**, *29*, 305–316. [CrossRef]
7. Vassalli, A.; Dellepiane, J.M.; Emmenegger, Y.; Jimenez, S.; Vandi, S.; Plazzi, G.; Franken, P.; Tafti, M. Electroencephalogram paroxysmal theta characterizes cataplexy in mice and children. *Brain* **2013**, *136*, 1592–1608. [CrossRef]
8. Paterson, L.M.; Nutt, D.J.; Wilson, S.J. Sleep and its disorders in translational medicine. *J. Psychopharmacol.* **2011**, *25*, 1226–1234. [CrossRef]
9. Lacroix, M.M.; de Lavilléon, G.; Lefort, J.; El Kanbi, K.; Bagur, S.; Laventure, S.; Dauvilliers, Y.; Peyron, C.; Benchenane, K. Improved sleep scoring in mice reveals human-like stages. *bioRxiv* 2018. [CrossRef]
10. Penzel, T.; Conradt, R. Computer based sleep recording and analysis. *Sleep Med. Rev.* **2000**, *4*, 131–148. [CrossRef]
11. Crisler, S.; Morrissey, M.J.; Anch, A.M.; Barnett, D.W. Sleep-stage scoring in the rat using a support vector machine. *J. Neurosci. Methods* **2008**, *168*, 524–534. [CrossRef]
12. Gao, V.; Turek, F.; Vitaterna, M. Multiple classifier systems for automatic sleep scoring in mice. *J. Neurosci. Methods* **2016**, *264*, 33–39. [CrossRef]
13. Svetnik, V.; Wang, T.C.; Xu, Y.; Hansen, B.J.; Fox, S.V. A Deep Learning Approach for Automated Sleep-Wake Scoring in Pre-Clinical Animal Models. *J. Neurosci. Methods* **2020**, *337*, 108668. [CrossRef]
14. Chen, T.; Guestrin, C. XGBoost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
15. Brankack, J.; Kukushka, V.I.; Vyssotski, A.L.; Draguhn, A. EEG gamma frequency and sleep-wake scoring in mice: Comparing two types of supervised classifiers. *Brain Res.* **2010**, *1322*, 59–71. [CrossRef] [PubMed]
16. Shao, X.; Soo Kim, C. A Hybrid Deep Learning Scheme for Multi-Channel Sleep Stage Classification. *Comput. Mater. Contin.* **2022**, *71*, 889–905. [CrossRef]
17. Miladinovic, D.; Muheim, C.; Bauer, S.; Spinnler, A.; Noain, D.; Bandarabadi, M.; Gallusser, B.; Krummenacher, G.; Baumann, C.; Adamantidis, A.; et al. SPINDLE: End-to-end learning from EEG/EMG to extrapolate animal sleep scoring across experimental settings, labs and species. *PLoS Comput. Biol.* **2019**, *15*, e1006968. [CrossRef] [PubMed]
18. Malafeev, A.; Laptev, D.; Bauer, S.; Omlin, X.; Wierzbicka, A.; Wichniak, A.; Jernajczyk, W.; Riener, R.; Buhmann, J.; Achermann, P. Automatic Human Sleep Stage Scoring Using Deep Neural Networks. *Front. Neurosci.* **2018**, *12*, 781. [CrossRef] [PubMed]
19. Bresch, E.; Grossekathofer, U.; Garcia-Molina, G. Recurrent Deep Neural Networks for Real-Time Sleep Stage Classification from Single Channel EEG. *Front. Comput. Neurosci.* **2018**, *12*, 85. [CrossRef] [PubMed]
20. Yue, H.; Chen, Z.; Guo, W.; Sun, L.; Dai, Y.; Wang, Y.; Ma, W.; Fan, X.; Wen, W.; Lei, W. Research and application of deep learning-based sleep staging: Data, modeling, validation, and clinical practice. *Sleep Med. Rev.* **2024**, *74*, 101897. [CrossRef]
21. Fiorillo, L.; Puiatti, A.; Papandrea, M.; Ratti, P.L.; Favaro, P.; Roth, C.; Bargiotas, P.; Bassetti, C.L.; Faraci, F.D. Automated sleep scoring: A review of the latest approaches. *Sleep Med. Rev.* **2019**, *48*, 101204. [CrossRef]
22. Sultana, A.; Islam, R. Machine learning framework with feature selection approaches for thyroid disease classification and associated risk factors identification. *J. Electr. Syst. Inf. Technol.* **2023**, *10*, 1–23. [CrossRef]
23. Bastianini, S.; Berteotti, C.; Gabrielli, A.; Lo Martire, V.; Silvani, A.; Zoccoli, G. Recent developments in automatic scoring of rodent sleep. *Arch. Ital. Biol.* **2015**, *153*, 58–66.
24. Avci, C.; Akbas, A. Sleep apnea classification based on respiration signals by using ensemble methods. *Biomed. Mater. Eng.* **2015**, *26* (Suppl. S1), S1703–S1710.
25. Sun, J.; Hu, X.; Peng, S.; Peng, C.K.; Ma, Y. Automatic classification of excitation location of snoring sounds. *J. Clin. Sleep Med.* **2021**, *17*, 1031–1038. [CrossRef]
26. Lazazzera, R.; Deviaene, M.; Varon, C.; Buyse, B.; Testelmans, D.; Laguna, P.; Gil, E.; Carrault, G. Detection and Classification of Sleep Apnea and Hypopnea Using PPG and SpO (2) Signals. *IEEE Trans. Biomed. Eng.* **2021**, *68*, 1496–1506. [CrossRef]

27. Toma, T.I.; Choi, S. An End-to-End Multi-Channel Convolutional Bi-LSTM Network for Automatic Sleep Stage Detection. *Sensors* **2023**, *23*, 4950. [CrossRef]

28. Sun, C.; Chen, C.; Fan, J.; Li, W.; Zhang, Y.; Chen, W. A hierarchical sequential neural network with feature fusion for sleep staging based on EOG and RR signals. *J. Neural. Eng.* **2019**, *16*, 066020. [CrossRef]

29. Barger, Z.; Frye, C.G.; Liu, D.; Dan, Y.; Bouchard, K.E. Robust, automated sleep scoring by a compact neural network with distributional shift correction. *PLoS ONE* **2019**, *14*, e0224642. [CrossRef]

30. Tezuka, T.; Kumar, D.; Singh, S.; Koyanagi, I.; Naoi, T.; Sakaguchi, M. Real-time, automatic, open-source sleep stage classification system using single EEG for mice. *Sci. Rep.* **2021**, *11*, 11151. [CrossRef]

31. Aboalayon, K.; Faezipour, M.; Almuhammadi, W.; Moslehpour, S. Sleep Stage Classification Using EEG Signal Analysis: A Comprehensive Survey and New Investigation. *Entropy* **2016**, *18*, e18090272. [CrossRef]

32. Koch, H.; Christensen, J.A.; Frandsen, R.; Zoetmulder, M.; Arvastson, L.; Christensen, S.R.; Jennum, P.; Sorensen, H.B. Automatic sleep classification using a data-driven topic model reveals latent sleep states. *J. Neurosci. Methods* **2014**, *235*, 130–137. [CrossRef]

33. Boostani, R.; Karimzadeh, F.; Nami, M. A comparative review on sleep stage classification methods in patients and healthy individuals. *Comput. Methods Programs Biomed.* **2017**, *140*, 77–91. [CrossRef]

34. Rahman, M.M.; Bhuiyan, M.I.H.; Hassan, A.R. Sleep stage classification using single-channel EOG. *Comput. Biol. Med.* **2018**, *102*, 211–220. [CrossRef]

35. Yoon, Y.S.; Hahm, J.; Kim, K.K.; Park, S.K.; Oh, S.W. Non-contact home-adapted device estimates sleep stages in middle-aged men: A preliminary study. *Technol. Health Care* **2020**, *28*, 439–446. [CrossRef]

36. Phan, H.; Andreotti, F.; Cooray, N.; Chen, O.Y.; De Vos, M. Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification. *IEEE Trans. Biomed. Eng.* **2019**, *66*, 1285–1296. [CrossRef] [PubMed]

37. Chambon, S.; Galtier, M.N.; Arnal, P.J.; Wainrib, G.; Gramfort, A. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 758–769. [CrossRef] [PubMed]

38. Phan, H.; Andreotti, F.; Cooray, N.; Chen, O.Y.; Vos, M. Automatic Sleep Stage Classification Using Single-Channel EEG: Learning Sequential Features with Attention-Based Recurrent Neural Networks. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2018**, *2018*, 1452–1455.

39. Fraiwan, L.; Alkhodari, M. Neonatal sleep stage identification using long short-term memory learning system. *Med. Biol. Eng. Comput.* **2020**, *58*, 1383–1391. [CrossRef] [PubMed]

40. Dong, H.; Supratak, A.; Pan, W.; Wu, C.; Matthews, P.M.; Guo, Y. Mixed Neural Network Approach for Temporal Sleep Stage Classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 324–333. [CrossRef]

41. Stephansen, J.B.; Olesen, A.N.; Olsen, M.; Ambati, A.; Leary, E.B.; Moore, H.E.; Carrillo, O.; Lin, L.; Han, F.; Yan, H.; et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat. Commun.* **2018**, *9*, 5229. [CrossRef]

42. Einizade, A.; Nasiri, S.; Sardouie, S.H.; Clifford, G.D. ProductGraphSleepNet: Sleep staging using product spatio-temporal graph learning with attentive temporal aggregation. *Neural Netw.* **2023**, *164*, 667–680. [CrossRef]

43. Awais, M.; Long, X.; Yin, B.; Farooq Abbasi, S.; Akbarzadeh, S.; Lu, C.; Wang, X.; Wang, L.; Zhang, J.; Dudink, J.; et al. A Hybrid DCNN-SVM Model for Classifying Neonatal Sleep and Wake States Based on Facial Expressions in Video. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 1441–1449. [CrossRef] [PubMed]

44. Chen, Z.; Wu, M.; Cui, W.; Liu, C.; Li, X. An Attention Based CNN-LSTM Approach for Sleep-Wake Detection with Heterogeneous Sensors. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3270–3277. [CrossRef] [PubMed]

45. Lajnef, T.; Chaibi, S.; Ruby, P.; Aguera, P.E.; Eichenlaub, J.B.; Samet, M.; Kachouri, A.; Jerbi, K. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. *J. Neurosci. Methods* **2015**, *250*, 94–105. [CrossRef] [PubMed]

46. Motin, M.A.; Karmakar, C.; Palaniswami, M.; Penzel, T.; Kumar, D. Multi-stage sleep classification using photoplethysmographic sensor. *R Soc. Open Sci.* **2023**, *10*, 221517. [CrossRef]

47. Zaman, A.; Kumar, S.; Shatabda, S.; Dehzangi, I.; Sharma, A. SleepBoost: A multi-level tree-based ensemble model for automatic sleep stage classification. *Med. Biol. Eng. Comput.* **2024**, *62*, 2769–2783. [CrossRef] [PubMed]

48. Tripathy, R.K.; Rajendra Acharya, U. Use of features from RR-time series and EEG signals for automated classification of sleep stages in deep neural network framework. *Biocybern. Biomed. Eng.* **2018**, *38*, 890–902. [CrossRef]

49. 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]

50. Abbasi, S.F.; Ahmad, J.; Tahir, A.; Awais, M.; Chen, C.; Irfan, M.; Siddiqa, H.A.; Waqas, A.B.; Long, X.; Yin, B.; et al. EEG-Based Neonatal Sleep-Wake Classification Using Multilayer Perceptron Neural Network. *IEEE Access* **2020**, *8*, 183025–183034. [CrossRef]

51. Ansari, A.H.; De Wel, O.; Pillay, K.; Dereymaeker, A.; Jansen, K.; Van Huffel, S.; Naulaers, G.; De Vos, M. A convolutional neural network outperforming state-of-the-art sleep staging algorithms for both preterm and term infants. *J. Neural Eng.* **2020**, *17*, 016028. [CrossRef] [PubMed]

52. Shao, Y.; Huang, B.; Du, L.; Wang, P.; Li, Z.; Liu, Z.; Zhou, L.; Song, Y.; Chen, X.; Fang, Z. Reliable automatic sleep stage classification based on hybrid intelligence. *Comput. Biol. Med.* **2024**, *173*, 108314. [CrossRef]

53. Hunt, J.; Coulson, E.J.; Rajnarayanan, R.; Oster, H.; Videnovic, A.; Rawashdeh, O. Sleep and circadian rhythms in Parkinson's disease and preclinical models. *Mol. Neurodegener.* **2022**, *17*, 2. [CrossRef] [PubMed]

54. Fraigne, J.J.; Wang, J.; Lee, H.; Luke, R.; Pintwala, S.K.; Peever, J.H. A novel machine learning system for identifying sleep-wake states in mice. *Sleep* **2023**, *46*, zsad101. [CrossRef] [PubMed]

55. Gross, B.A.; Walsh, C.M.; Turakhia, A.A.; Booth, V.; Mashour, G.A.; Poe, G.R. Open-source logic-based automated sleep scoring software using electrophysiological recordings in rats. *J. Neurosci. Methods* **2009**, *184*, 10–18. [CrossRef] [PubMed]

56. Rempe, M.J.; Clegern, W.C.; Wisor, J.P. An automated sleep-state classification algorithm for quantifying sleep timing and sleep-dependent dynamics of electroencephalographic and cerebral metabolic parameters. *Nat. Sci. Sleep* **2015**, *7*, 85–99. [CrossRef]

57. Exarchos, I.; Rogers, A.A.; Aiani, L.M.; Gross, R.E.; Clifford, G.D.; Pedersen, N.P.; Willie, J.T. Supervised and unsupervised machine learning for automated scoring of sleep-wake and cataplexy in a mouse model of narcolepsy. *Sleep* **2020**, *43*, zsz272. [CrossRef]

58. Yamabe, M.; Horie, K.; Shiokawa, H.; Funato, H.; Yanagisawa, M.; Kitagawa, H. MC-SleepNet: Large-scale Sleep Stage Scoring in Mice by Deep Neural Networks. *Sci. Rep.* **2019**, *9*, 15793. [CrossRef]

59. Jha, P.K.; Valekunja, U.K.; Reddy, A.B. SlumberNet: Deep learning classification of sleep stages using residual neural networks. *Sci. Rep.* **2024**, *14*, 4797. [CrossRef]

60. Kam, K.; Rapoport, D.M.; Parekh, A.; Ayappa, I.; Varga, A.W. WaveSleepNet: An interpretable deep convolutional neural network for the continuous classification of mouse sleep and wake. *J. Neurosci. Methods* **2021**, *360*, 109224. [CrossRef]

61. Meng, Q.; Tan, X.; Jiang, C.; Xiong, Y.; Yan, B.; Zhang, J. Tracking Eye Movements During Sleep in Mice. *Front. Neurosci.* **2021**, *15*, 616760. [CrossRef]

62. Senzai, Y.; Scanziani, M. A cognitive process occurring during sleep is revealed by rapid eye movements. *Science* **2022**, *377*, 999–1004. [CrossRef]

63. Judge, S.J.; Richmond, B.J.; Chu, F.C. Implantation of magnetic search coils for measurement of eye position: An improved method. *Vis. Res.* **1980**, *20*, 535–538. [CrossRef]

64. Pei, W.; Li, Y.; Wen, P.; Yang, F.; Ji, X. An automatic method using MFCC features for sleep stage classification. *Brain Inform.* **2024**, *11*, 6. [CrossRef]

65. Fan, J.; Sun, C.; Chen, C.; Jiang, X.; Liu, X.; Zhao, X.; Meng, L.; Dai, C.; Chen, W. EEG data augmentation: Towards class imbalance problem in sleep staging tasks. *J. Neural Eng.* **2020**, *17*, 056017. [CrossRef]

66. Grieger, N.; Schwabedal, J.T.C.; Wendel, S.; Ritze, Y.; Bialonski, S. Automated scoring of pre-REM sleep in mice with deep learning. *Sci. Rep.* **2021**, *11*, 12245. [CrossRef]

67. Tsinalis, O.; Matthews, P.M.; Guo, Y. Automatic Sleep Stage Scoring Using Time-Frequency Analysis and Stacked Sparse Autoencoders. *Ann. Biomed. Eng.* **2016**, *44*, 1587–1597. [CrossRef]

68. Sunagawa, G.A.; Sei, H.; Shimba, S.; Urade, Y.; Ueda, H.R. FASTER: An unsupervised fully automated sleep staging method for mice. *Genes Cells* **2013**, *18*, 502–518. [CrossRef]

69. Aserinsky, E.; Kleitman, N. Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. *Science* **1953**, *118*, 273–274. [CrossRef]

70. Andrillon, T.; Nir, Y.; Cirelli, C.; Tononi, G.; Fried, I. Single-neuron activity and eye movements during human REM sleep and awake vision. *Nat. Commun.* **2015**, *6*, 7884. [CrossRef]

71. Fulda, S.; Romanowski, C.P.; Becker, A.; Wetter, T.C.; Kimura, M.; Fenzel, T. Rapid eye movements during sleep in mice: High trait-like stability qualifies rapid eye movement density for characterization of phenotypic variation in sleep patterns of rodents. *BMC Neurosci.* **2011**, *12*, 110. [CrossRef]

72. Ronzhina, M.; Janousek, O.; Kolarova, J.; Novakova, M.; Honzik, P.; Provaznik, I. Sleep scoring using artificial neural networks. *Sleep Med. Rev.* **2012**, *16*, 251–263. [CrossRef]

73. Yin, J.; Xu, J.; Ren, T.L. Recent Progress in Long-Term Sleep Monitoring Technology. *Biosensors* **2023**, *13*, 395. [CrossRef]