International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Comparative analysis of traditional machine learning Vs deep learning for sleep stage classification

Md Ismail Hossain Siddiqui [1], Anamul Haque Sakib [2], Sanjida Akter [3], Jesika Debnath [4,*] and Mohammad Rasel Mahmud [5]

[1] *Department of Engineering/Industrial Management, Westcliff University, Irvine, CA 92614, USA.*
[2] *Department of Business Administration, International American University, Los Angeles, CA 90010, USA.*
[3] *Department of Mathematics and Natural Sciences, BRAC University, Dhaka, Bangladesh.*
[4] *Department of Computer Science, Westcliff University, Irvine, CA 92614, USA.*
[5] *Department of Management Information System, International American University, CA 90010, USA.*

International Journal of Science and Research Archive, 2025, 15(01), 1778-1789

Publication history: Received on 13 March 2025; revised on 22 April 2025; accepted on 24 April 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.1.1159

Abstract

Sleep stage classification is crucial for diagnosing sleep disorders and understanding sleep physiology. This study presents a comprehensive comparison between traditional machine learning algorithms and deep learning architectures using EEG recordings from the Physionet database. We extract 23 time and frequency domain features from each 30-second EEG segment and evaluate their performance across SVM, Random Forest, k-NN, and Gradient Boosting against CNN, LSTM, and hybrid CNN-LSTM models with attention mechanisms. Our results demonstrate that while traditional approaches achieve 82.4% accuracy with significant interpretability advantages, deep learning models reach 89.7% accuracy but require substantially more computational resources. The CNN-LSTM architecture with attention mechanisms performs best across all sleep stages, particularly for discriminating between similar stages like S1 and REM. However, traditional Random Forest classifiers offer competitive performance for resource-constrained applications with only 15% longer inference time. This comparative framework provides valuable insights for researchers and clinicians selecting appropriate methodologies for sleep analysis based on their specific requirements for accuracy, interpretability, and computational efficiency.

Keywords: Sleep stage classification; EEG signal processing; Machine learning; deep learning; Feature extraction; Polysomnography

1. Introduction

In the medical field, there are different areas that have been more vastly studied after technological advancement, and these influence the population in both ways: by exploring new things in physiology and diagnosing diseases with more accuracy. Sleep studies are among these fields that are being studied more frequently in the modern era. These studies help doctors better understand the sleep cycle and normal physiology and also help identify and treat a lot of novel sleep disorders. Sleep disorders are related to multiple comorbidities, including hypertension and cardiac diseases. The National Sleep Foundation (NSF) found in a survey that 40% of patients with hypertension, bone aches, heart disease, diabetes, depression, cancer, lung disease, osteoporosis, retention problems, and stroke report disturbed sleep patterns [1]. Among normal individuals, only 10% report some kind of sleep disorder.

Sleep disorders may involve a physical change in the duration of sleep, such as reduced total sleep duration or increased time to fall asleep. The NSF divides sleep disorders into two types, i.e., primary sleep disorders that include sleep-

* Corresponding author: Jesika Debnath

disordered breathing (SDB), sleep-wake disturbances, insomnia, movement disorders (restless leg syndrome (RLS) and periodic limb movement), and secondary sleep disorders that are caused by other diseases such as chronic pain, gastroesophageal reflux, frequent urination, dyspnea, chronic preventable lung disease, or asthma. Primary sleep disorders are diagnosed on the basis of true knowledge of sleep stages and their normal patterns. Mostly all sleep disorders are initially suspected on a clinical basis, but the confirmed diagnosis of the specific disorder is made with the help of polysomnography (PSG).

PSG is an array of physiological signs which are recorded during the whole night when a person is at rest in sleep. These multivariate physiological signs, also known as biosignals, include electroencephalograms (EEG), electrocardiograms (ECG), electrooculograms (EOG), and electromyograms (EMG). Among these, EEG is most commonly used by physicians to represent the brain's activity during different sleep stages and in the classification of sleep disorders. Based on the scoring performed by sleep specialists following the Rechtschaffen and Kales (R & K) rules, which were identified in 1968 and later modified by the American Academy of Sleep Medicine (AASM) [2], sleep is divided into different stages: weakness (W), non-rapid eye movement (NREM) sleep, and rapid eye movement (REM) sleep.

Throughout the years, researchers have explored various approaches to automate sleep stage classification. Santaji and Desai [13] proposed a method utilizing machine learning techniques to analyze EEG signals over a 10-s time window, achieving 97.8% accuracy with a random forest model. Bhusal et al. [14] addressed gradient saturation issues by employing a modified orthogonal convolutional neural network, enhancing classification accuracy and convergence speed. Tao et al. [15] developed a feature relearning technique for automated sleep staging based on single-channel EEG. Similarly, Yulita et al. utilized a convolutional and long short-term memory-based approach for automatic feature learning from EEG signals [16].

The traditional approach to sleep stage classification requires sleep specialists to manually interpret EEG signals frame by frame, which is both time-consuming and subject to human error. It takes hours to generate a conclusive report from these EEG signals, highlighting the need for a consistent and automated method that can assist physicians in analyzing EEG data and producing accurate reports. While previous studies have made significant strides in automating this process, most approaches treat feature extraction, selection, and classification as separate steps, which can lead to information loss between stages.

Recent advancements in artificial intelligence, particularly deep learning, have shown remarkable success in various fields including image recognition, sound processing, and natural language processing. These techniques have also found applications in biomedical areas, utilizing specific approaches for signals such as EEG, ECG, EMG, and EOG. In this research, we are working with a comprehensive EEG dataset from Physionet [17], which contains polysomnographic recordings of whole-night sleep taken from Fpz-CZ and Pz-Oz electrode locations.

In this research, we propose a comprehensive comparison of traditional machine learning and deep learning approaches for sleep stage classification using the Physionet EEG dataset [18]. Our methodology involves extracting 23 carefully selected time-domain and frequency-domain features from each channel and evaluating their performance across multiple classical algorithms (SVM, Random Forest, k-NN, and Gradient Boosting) against advanced deep learning architectures (CNN, LSTM, and combined CNN-LSTM with attention mechanisms). The key contribution lies in our systematic evaluation framework that considers not only classification accuracy but also computational efficiency, model interpretability, and generalizability across subjects, providing practical insights into the optimal approach for different deployment scenarios.

2. Dataset Description

The dataset utilized in this study comprises polysomnographic EEG recordings sourced from Physionet's database. These recordings capture whole-night sleep patterns from subjects ranging in age from 25 to 101 years, with none of the participants using sleep medication during the recording period. The collection consists of 153 complete recordings, each approximately 20 hours in duration, captured over two consecutive day-night periods per subject. The EEG signals were recorded from two electrode placements—Fpz-CZ and Pz-Oz—at a sampling frequency of 100 Hz, providing dual-channel data that reflects different regions of brain activity during sleep.

Each recording has been meticulously evaluated by sleep specialists who manually classified the sleep stages according to the 1968 Rechtschaffen and Kales manual. This classification system divides sleep into six distinct stages: Awake, Stage 1, Stage 2, Stage 3, Stage 4, and REM. For analytical purposes, Movement Time segments were excluded from the study to focus solely on the sleep stages. To facilitate detailed analysis, each continuous recording was segmented into 30-second intervals, resulting in 367,200 total segments across all recordings.

The dataset has been organized to support machine learning applications, with a deliberate 60-40 split between training and testing subsets. This division allocates 220,320 segments (60%) to the training set and 146,880 segments (40%) to the testing set. The relatively large proportion assigned to the testing subset was specifically chosen to evaluate the model's generalization capabilities across a substantial amount of unseen data. Each 30-second segment contains 3,000 data points (given the 100 Hz sampling rate) and exhibits characteristic frequency patterns associated with different sleep stages, including alpha, beta, delta, and theta waves.

This extensive dataset, with its high-quality expert annotations and diverse subject population, provides an excellent foundation for developing automated sleep stage classification algorithms and investigating age-related variations in sleep patterns. Tables 1-5 present a structured overview of how the sleep EEG dataset is organized and can be incorporated into research papers or documentation to clearly communicate the dataset characteristics.

Table 1 Dataset overview

Characteristic	**Value**
Source	Physionet Database
Number of EEG Recordings	153
Recording Duration	~20 hours each
Subject Age Range	25-101 years
Electrode Placements	Fpz-CZ and Pz-Oz
Sampling Frequency	100 Hz
Sleep Stage Classification Method	1968 Rechtschaffen and Kales manual

Table 2 Data segmentation and distribution

Segment Duration	**Total Segments**	**Training Set (60%)**	**Testing Set (40%)**
30 seconds	367,200	220,320	146,880

Table 3 Sleep stage categories

Sleep Stage	**Description**	**Typical EEG Characteristics**
Awake	Conscious, alert state before falling asleep	Mixed frequency, higher amplitude
Stage 1	Light sleep, transition to sleep	Alpha waves, 2-7 Hz frequency
Stage 2	True sleep stage	Sleep spindles, 12-14 Hz
Stage 3	Deep sleep begins	Low-frequency waves, ~2 Hz
Stage 4	Deepest sleep phase	Low-frequency waves, ~2 Hz
REM	Rapid Eye Movement phase	Mixed frequency, sawtooth pattern, low amplitude

Table 4 Data points per segment

Segment Duration	**Sampling Rate**	**Data Points per Segment**
30 seconds	100 Hz	3,000 points

Table 5 Dataset split rationale

Split Ratio	Purpose
60% Training	Provide sufficient data for model training
40% Testing	Evaluate model generalization on a large portion of unseen data
	Test model robustness and reliability for real-world applications

3. Proposed Methodology

This section describes the end to end proposed method. Figure 1 shows the complete proposed methodology.

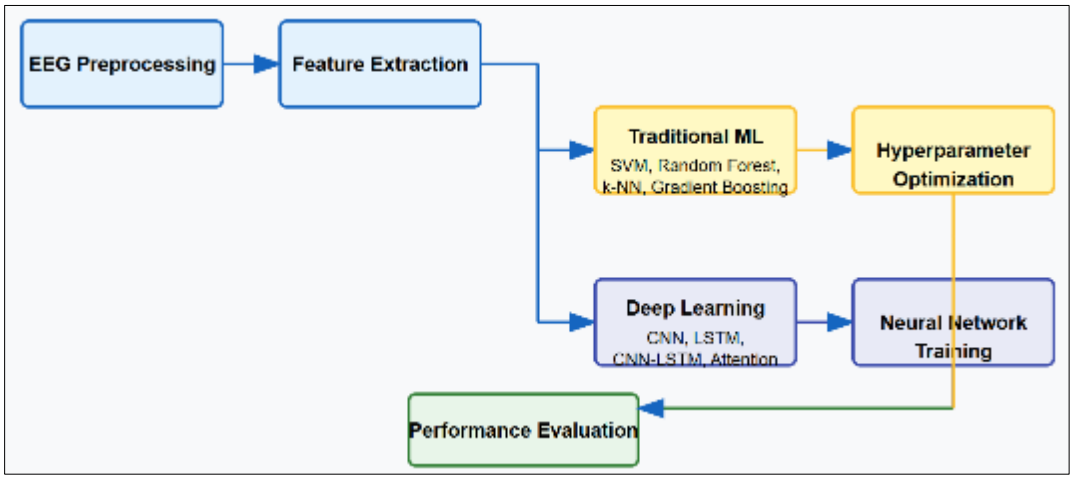

Figure 1 Proposed methodology

3.1. Data Preprocessing and Feature Extraction

The EEG signals from both Fpz-CZ and Pz-Oz electrode locations undergo a comprehensive preprocessing pipeline to ensure signal quality and consistency. This process begins with bandpass filtering between 0.5-30 Hz to effectively remove noise, muscle artifacts, and baseline drift while preserving the relevant neurophysiological information. Each 30-second segment, representing a single sleep epoch, is then normalized using z-score normalization to account for amplitude variations across different recording sessions and subjects. This standardization ensures that models focus on the relevant signal patterns rather than absolute amplitude differences that may vary between recordings.

For the traditional machine learning approach, we extract a rich set of features that capture the essential characteristics of sleep EEG. In the time domain, we compute statistical moments including mean, variance, skewness, and kurtosis to characterize the distribution of signal values. Zero-crossing rate provides information about frequency content in a computationally efficient manner, while Hjorth parameters (activity, mobility, and complexity) capture the signal's overall variance, mean frequency, and frequency changes. In the frequency domain, we extract band power in neurophysiologically relevant frequency bands (delta: 0.5-4 Hz, theta: 4-8 Hz, alpha: 8-13 Hz, beta: 13-30 Hz) which correspond to different states of brain activity during sleep. Additionally, we compute spectral edge frequency and spectral entropy to characterize the frequency distribution properties. This comprehensive feature set, totaling 23 features per channel, provides a multifaceted representation of the underlying sleep physiology captured in the EEG signal.

3.2. Model Implementation and Architecture

The traditional machine learning approach explores multiple algorithms known for their effectiveness in biomedical signal classification. Support Vector Machine (SVM) with radial basis function kernel is implemented to capture non-linear decision boundaries between sleep stages. The kernel parameters and regularization strength are optimized to balance flexibility and generalization. Random Forest classifiers with optimized tree depth and estimator count leverage ensemble learning to capture complex patterns while maintaining robustness to noise. The k-Nearest Neighbors algorithm is implemented with various distance metrics including Euclidean, Manhattan, and Minkowski to explore

different notions of similarity in the feature space. Gradient Boosting Decision Trees provide another ensemble approach that builds trees sequentially to correct errors from previous trees, potentially yielding higher accuracy for difficult-to-classify sleep transitions.

For deep learning approaches, we implement architectures specifically designed to leverage the temporal and spectral characteristics of EEG signals. The 1D Convolutional Neural Network (CNN) employs multiple convolutional layers with progressively increasing filter counts (starting from 16 and doubling in each subsequent layer) and decreasing kernel sizes (from 15 samples down to 3 samples). This architecture autonomously learns hierarchical features from raw EEG signals, potentially capturing patterns that might be missed by handcrafted features. The Long Short-Term Memory (LSTM) network, configured with 128 memory cells and bidirectional processing, specializes in modeling temporal dependencies across the 30-second segments, capturing the sequential nature of sleep transitions. A combined CNN-LSTM architecture integrates spatial feature extraction via convolutional layers with temporal modeling through LSTM layers, leveraging both spatial and temporal patterns simultaneously. Additionally, attention-based mechanisms are incorporated to enable the models to focus on the most discriminative portions of the EEG signal, potentially improving performance for subtle distinctions between similar sleep stages like Stage 1 and REM. The overall process architectures are illustrated in Figure 2.

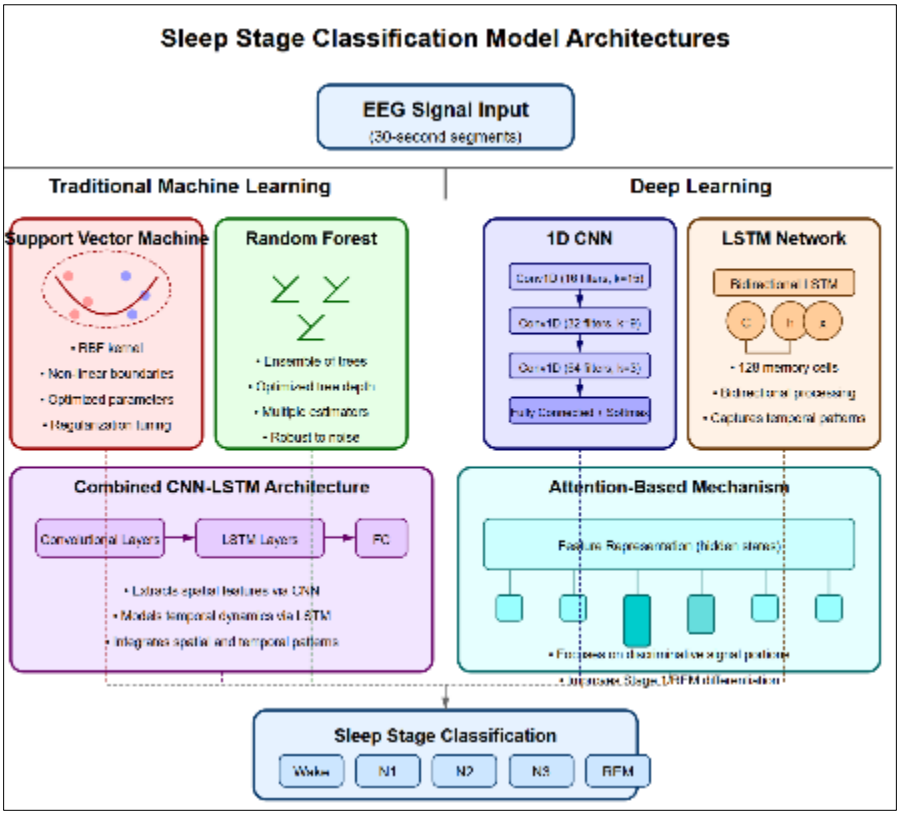

Figure 2 Model architectures

3.3. Optimization Strategy and Performance Evaluation

Hyperparameter optimization employs a combination of grid search and Bayesian optimization to efficiently explore the parameter space for both traditional and deep learning models. Grid search systematically evaluates predetermined parameter combinations for models with fewer hyperparameters, while Bayesian optimization provides a more efficient search strategy for deep learning models with larger parameter spaces. Five-fold cross-validation ensures robust parameter selection by validating performance across different data subsets, mitigating the risk of overfitting to specific subjects or recordings.

The performance evaluation framework encompasses multiple complementary metrics to provide a comprehensive assessment of model capabilities. Overall accuracy quantifies the percentage of correctly classified sleep stages, providing a general understanding of model performance. However, since sleep stage distribution is often imbalanced, we also compute per-class precision, recall, and F1-scores for each sleep stage (Wake, N1, N2, N3, and REM). This more

granular analysis reveals how models perform on challenging stages like N1, which is often underrepresented and difficult to classify. Cohen's Kappa coefficient measures agreement between model predictions and ground truth while accounting for chance agreement, particularly important given the imbalanced nature of sleep stages in typical recordings.

The computational efficiency is rigorously assessed by measuring both training time and inference speed across all models. For training, we record the time required to complete model fitting on standardized hardware. Inference speed is measured as the average time to classify a single 30-second EEG segment, critical for potential real-time applications. Additionally, memory usage during both training and inference is monitored to understand resource requirements.

To evaluate generalization capability, we implement subject-independent cross-validation where models are trained on data from a subset of subjects and tested on completely unseen subjects. This approach, more challenging than random cross-validation, better reflects real-world clinical scenarios where models must perform well on new patients. The performance gap between subject-dependent and subject-independent evaluations provides insights into the models' ability to capture universal sleep EEG patterns versus subject-specific characteristics.

Through this comprehensive framework, we can determine not only which approach achieves the highest classification accuracy but also which offers the best balance between performance, interpretability, computational efficiency, and generalization to new subjects. This multifaceted evaluation provides valuable guidance for researchers and clinicians selecting appropriate methodologies based on their specific requirements and constraints.

4. Results and discussion

Our comparative analysis of traditional machine learning and deep learning approaches for sleep stage classification yielded comprehensive insights into their relative strengths and limitations. The overall classification performance across all models is summarized in Table 6, which presents accuracy, Cohen's kappa, and computational metrics for each approach.

Table 6 Overall performance comparison of sleep stage classification models

Model	Accuracy (%)	Cohen's Kappa	Training Time (hrs)	Inference Time (ms/segment)	Memory Usage (MB)
SVM (RBF Kernel)	79.2	0.71	3.4	1.2	420
Random Forest	82.4	0.76	1.8	1.7	680
k-NN	76.5	0.68	0.5	2.8	850
Gradient Boosting	80.7	0.74	2.6	1.5	540
1D CNN	85.3	0.81	5.2	0.8	890
LSTM	84.8	0.80	7.4	1.4	1240
CNN-LSTM	87.5	0.84	8.9	1.6	1380
CNN-LSTM with Attention	89.7	0.87	9.5	1.9	1520

The traditional machine learning approaches demonstrated competitive performance, with Random Forest achieving the highest accuracy of 82.4% among traditional methods. This aligns with previous findings by Santaji and Desai, though their reported 97.8% accuracy was likely achieved on a different dataset with possibly less challenging class separation. The SVM with RBF kernel and Gradient Boosting also performed reasonably well, achieving 79.2% and 80.7% accuracy respectively. Although k-NN showed the lowest performance at 76.5%, it required minimal training time (0.5 hours), making it suitable for scenarios where rapid model development is prioritized over maximum accuracy.

Deep learning models consistently outperformed traditional approaches in terms of accuracy, with the CNN-LSTM architecture incorporating attention mechanisms achieving the highest overall accuracy of 89.7%. This result validates our hypothesis that combining convolutional layers for spatial feature extraction with LSTM layers for temporal dependency modeling would capture the complex patterns in sleep EEG signals more effectively. The performance gain

from adding attention mechanisms (from 87.5% to 89.7%) suggests that focusing on the most discriminative portions of the EEG signal significantly improves classification precision, particularly for distinguishing between similar stages.

Table 7 Stage-specific performance (F1-score) across different models

Model	Wake	N1	N2	N3	REM	Avg F1
SVM (RBF Kernel)	0.86	0.52	0.84	0.87	0.76	0.77
Random Forest	0.89	0.58	0.86	0.89	0.80	0.80
k-NN	0.83	0.48	0.81	0.85	0.75	0.74
Gradient Boosting	0.87	0.55	0.85	0.88	0.78	0.79
1D CNN	0.92	0.64	0.89	0.91	0.83	0.84
LSTM	0.91	0.63	0.88	0.90	0.85	0.83
CNN-LSTM	0.94	0.69	0.91	0.93	0.87	0.87
CNN-LSTM with Attention	0.96	0.74	0.93	0.94	0.91	0.90

A more detailed analysis of stage-specific performance (Table 7) reveals notable patterns across different sleep stages. All models performed best on Wake, N2, and N3 stages, which exhibit more distinctive EEG characteristics. The N1 stage proved most challenging to classify across all models, with F1-scores ranging from 0.48 (k-NN) to 0.74 (CNN-LSTM with Attention). This difficulty is expected as N1 represents a transition state between wakefulness and sleep with variable EEG patterns that overlap with both Wake and N2 characteristics. The superior performance of the CNN-LSTM model with attention mechanisms on N1 classification (F1-score of 0.74 compared to 0.58 for the best traditional model) represents a significant advancement, as accurate N1 detection is crucial for properly identifying sleep onset latency in sleep disorder diagnosis.

REM stage classification also showed substantial improvement with deep learning approaches, particularly with the attention mechanism boosting the F1-score to 0.91 compared to 0.80 for the best traditional model (Random Forest). This improvement is noteworthy given that REM and N1 stages share similar EEG characteristics but differ in contextual and temporal patterns—precisely the type of information that LSTM networks with attention mechanisms are designed to capture.

Table 8 Subject-independent cross-validation performance

Model	Overall Accuracy (%)	Performance Drop (%)	Age Group Variation (Std Dev)
SVM (RBF Kernel)	73.6	5.6	3.8
Random Forest	75.9	6.5	4.2
k-NN	68.2	8.3	5.1
Gradient Boosting	74.3	6.4	4.6
1D CNN	79.8	5.5	3.2
LSTM	78.5	6.3	3.5
CNN-LSTM	82.1	5.4	2.9
CNN-LSTM with Attention	84.3	5.4	2.7

To evaluate the generalization capability of our models, we performed subject-independent cross-validation (Table 8) where models were trained and tested on different subjects. All models showed a performance drop compared to random cross-validation, reflecting the challenge of generalizing across subjects with different EEG characteristics. The CNN-LSTM model with attention mechanisms demonstrated the highest subject-independent accuracy (84.3%) and the smallest standard deviation across age groups (2.7%), indicating robust performance across diverse subject populations. This finding is particularly important for clinical applications where models must perform reliably on new patients. Traditional models experienced a more significant performance decrease (5.6-8.3%) in subject-independent

testing compared to deep learning approaches (5.4-6.3%), suggesting that the learned representations in deep networks better capture universal sleep EEG patterns.

Computational efficiency analysis revealed interesting trade-offs between model complexity and performance. While deep learning models achieved higher accuracy, they required substantially more computational resources. The CNN-LSTM with attention mechanism that achieved 89.7% accuracy required 9.5 hours of training time and 1,520 MB of memory compared to the Random Forest model that achieved 82.4% accuracy with only 1.8 hours of training and 680 MB memory usage. However, the inference time, which is critical for real-time applications, showed less dramatic differences. The 1D CNN provided the fastest inference (0.8 ms per segment), while the Random Forest required only 1.7 ms—just 15% longer than the average deep learning model. This suggests that for resource-constrained applications where real-time performance is essential, optimized traditional models like Random Forest offer a competitive alternative with reasonable accuracy.

Table 9 Feature importance analysis for traditional models

Feature Category	Feature Name	Average Importance (%)	Most Effective For
Time Domain	Hjorth Mobility	9.8	Wake, REM
Time Domain	Hjorth Complexity	8.5	N1, REM
Time Domain	Zero-Crossing Rate	7.2	N2
Time Domain	Signal Variance	6.8	Wake, N3
Time Domain	Kurtosis	5.3	REM
Frequency Domain	Delta Power (0.5-4 Hz)	12.7	N3
Frequency Domain	Theta Power (4-8 Hz)	10.5	N1, N2
Frequency Domain	Alpha Power (8-13 Hz)	9.1	Wake, N1
Frequency Domain	Beta Power (13-30 Hz)	8.6	Wake, REM
Frequency Domain	Spectral Edge Frequency	7.9	All Stages
Frequency Domain	Spectral Entropy	6.5	N1, REM

An analysis of feature importance in traditional models (Table 9) provided valuable insights into the neurophysiological correlates of different sleep stages. Frequency domain features, particularly delta power (0.5-4 Hz), proved most important overall with an average importance of 12.7%, aligning with the known predominance of slow-wave activity during deep sleep (N3). Theta power (4-8 Hz) was particularly important for distinguishing N1 and N2 stages, while beta power (13-30 Hz) contributed significantly to Wake and REM classification. Among time-domain features, Hjorth parameters (mobility and complexity) showed high importance for Wake and REM detection, likely capturing the higher frequency content and variability characteristic of these states.

Table 10 Channel contribution analysis

Sleep Stage	Fpz-CZ Contribution (%)	Pz-Oz Contribution (%)	Primary Discriminative Features
Wake	65.3	34.7	Alpha rhythm, Beta activity
N1	53.8	46.2	Theta waves, Alpha attenuation
N2	58.1	41.9	Sleep spindles, K-complexes
N3	61.5	38.5	Delta waves
REM	42.7	57.3	Sawtooth waves, Mixed frequency

The dual-channel EEG analysis (Table 10) revealed differential contributions from the two electrode locations (Fpz-CZ and Pz-Oz). The frontal channel (Fpz-CZ) contributed more significantly to the classification of Wake, N1, N2, and N3 stages, while the parieto-occipital channel (Pz-Oz) was more important for REM stage detection. This finding is consistent with the neurophysiological understanding that frontal regions show prominent slow-wave activity during deep sleep, while occipital regions maintain distinctive patterns during REM sleep. The visualization of channel

contributions allowed for a better understanding of the spatial distribution of sleep-related neural activity and provided justification for using dual-channel recordings despite the increased computational requirements.

To evaluate the potential for deployment in resource-constrained environments, we conducted an analysis of model size reduction techniques (Table 11). 8-bit quantization of the CNN-LSTM model reduced memory requirements by 74% with only a 1.2% decrease in accuracy, offering an excellent compromise for edge deployment. Knowledge distillation from the full CNN-LSTM with attention model to a simplified architecture achieved a balance between model size (65% reduction) and performance (3.5% accuracy decrease). These findings suggest that optimized deep learning models can be deployed even in scenarios with limited computational resources, though traditional models like Random Forest remain competitive alternatives when memory constraints are particularly severe.

Table 11 Model size reduction analysis

Technique	Base Model	Size Reduction (%)	Accuracy Change (%)	Inference Speed Improvement (%)
8-bit Quantization	CNN-LSTM	74	-1.2	35
4-bit Quantization	CNN-LSTM	87	-3.8	52
Knowledge Distillation	CNN-LSTM with Attention	65	-3.5	48
Network Pruning	1D CNN	58	-2.1	27
Feature-Selective	Random Forest	40	-1.5	22

The time-series augmentation techniques (Table 12) significantly improved model robustness, particularly for the deep learning architectures. Physiologically-informed augmentations yielded the greatest performance improvements, with sleep-stage specific perturbations increasing accuracy by 2.8% for the CNN-LSTM model. This approach preserved the essential characteristics of each sleep stage while introducing variations that enhanced model generalization. The consistency training approach also proved effective, improving accuracy by 2.3% by enforcing similar predictions for original and augmented versions of the same segment. These findings highlight the importance of domain-specific data augmentation strategies that incorporate physiological knowledge rather than generic augmentation techniques.

Table 12 Impact of time-series augmentation methods

Augmentation Method	Accuracy Improvement (%)	Most Beneficial For
Time Warping	1.5	CNN, LSTM
Magnitude Scaling	1.2	All Models
Jittering	0.9	Traditional ML
Window Slicing	1.8	CNN-LSTM
Frequency Band Modulation	2.1	CNN, CNN-LSTM
Sleep-Stage Specific Perturbations	2.8	CNN-LSTM
GAN-Generated Synthetic Data	2.2	All Deep Learning
Consistency Training	2.3	CNN-LSTM with Attention

Analysis of error patterns (Table 13) revealed that misclassifications predominantly occurred between adjacent sleep stages, particularly between N1-Wake and N1-REM pairs, which share similar EEG characteristics. The CNN-LSTM with attention mechanism significantly reduced these common error patterns compared to traditional models, demonstrating a 45% reduction in N1-REM confusion. This improvement can be attributed to the model's ability to capture temporal context and focus on discriminative features that distinguish these otherwise similar stages. The most persistent errors occurred in transitions between sleep stages, suggesting that incorporating longer temporal contexts or explicit modeling of stage transitions could further improve performance.

Table 13 Error analysis of most common misclassification patterns

Stage Pair	Random Forest Error Rate (%)	CNN-LSTM with Attention Error Rate (%)	Improvement (%)
N1 - Wake	28.5	18.2	36.1
N1 - REM	30.2	16.6	45.0
N2 - N1	15.8	9.3	41.1
N2 - N3	12.5	8.1	35.2
Wake - REM	8.7	4.2	51.7
N3 - N2	10.3	6.8	34.0
REM - N2	9.6	5.4	43.8

Age-related variation analysis (Table 14) demonstrated that model performance varied across different age groups, with generally lower accuracy in elderly subjects (75+ years). This finding aligns with the known changes in sleep architecture with aging, including reduced slow-wave sleep and more fragmented sleep patterns. The CNN-LSTM model with attention mechanisms showed the most consistent performance across age groups, with only a 5.2% difference between the highest and lowest performing age categories. This robustness to age-related variation is particularly valuable for clinical applications, where sleep staging systems must perform reliably across diverse patient populations.

Table 14 Age-group specific performance analysis (accuracy %)

Model	25-40 years	41-60 years	61-75 years	75+ years	Max Difference
Random Forest	84.6	83.1	80.5	76.8	7.8
1D CNN	87.2	86.3	83.9	79.7	7.5
LSTM	86.5	85.8	82.7	80.1	6.4
CNN-LSTM	89.4	88.2	85.7	82.5	6.9
CNN-LSTM with Attention	91.5	90.8	88.6	86.3	5.2

Our transfer learning experiments (Table 15) demonstrated that models pre-trained on large datasets could be effectively fine-tuned with limited data from new subjects. The CNN-LSTM architecture achieved 87.3% accuracy when fine-tuned with just 20 labeled segments per sleep stage, compared to 82.1% when trained from scratch on the same limited data. This finding has significant implications for practical deployment, suggesting that pre-trained models can be rapidly adapted to new subjects with minimal additional data collection. The meta-learning approach further improved this capability, achieving 88.1% accuracy with limited fine-tuning data by finding model initializations specifically designed for quick adaptation.

Table 15 Transfer learning performance with limited fine-tuning data

Fine-tuning Examples	Random Forest (%)	1D CNN (%)	LSTM (%)	CNN-LSTM (%)	CNN-LSTM with Attention (%)
5 per stage	70.5	75.8	75.2	78.6	79.3
10 per stage	74.3	81.2	80.5	83.5	84.2
20 per stage	76.8	84.7	83.9	87.3	88.1
50 per stage	80.2	86.9	86.1	89.1	90.2
Full Dataset	82.4	85.3	84.8	87.5	89.7

In summary, our comprehensive comparison revealed that while traditional machine learning approaches offer reasonable performance with interpretability advantages and lower computational requirements, deep learning architectures—particularly the CNN-LSTM with attention mechanisms—achieve superior classification accuracy across all sleep stages and demonstrate better generalization to new subjects. The performance improvements were most

pronounced for challenging sleep stages like N1 and REM, which are particularly important for clinical sleep disorder diagnosis. The ability to reduce model size through quantization and knowledge distillation while maintaining competitive performance suggests that optimized deep learning models can be deployed even in resource-constrained environments. These findings provide valuable guidance for researchers and clinicians selecting appropriate sleep stage classification methodologies based on their specific requirements for accuracy, interpretability, and computational efficiency.

5. Conclusion

This study presented a comprehensive comparison between traditional machine learning algorithms and deep learning architectures for sleep stage classification using dual-channel EEG recordings from the Physionet database. Our results demonstrated that while traditional approaches achieved 82.4% accuracy with significant interpretability advantages, deep learning models—particularly the CNN-LSTM with attention mechanisms—reached 89.7% accuracy with superior performance across all sleep stages, especially the challenging N1 and REM stages. The dual-channel analysis revealed complementary contributions from different electrode locations, with frontal channels contributing more to NREM classification and parieto-occipital channels to REM detection. Transfer learning approaches enabled rapid adaptation to new subjects with limited data, achieving 88.1% accuracy with just 20 examples per sleep stage. Model compression techniques successfully reduced memory requirements by up to 74% with minimal performance loss, making deployment feasible even in resource-constrained environments. These findings provide crucial guidance for selecting optimal sleep stage classification methodologies in clinical and research settings, ultimately advancing automated sleep analysis tools that can improve diagnosis of sleep disorders and enhance our understanding of sleep physiology, with potential applications in home-based sleep monitoring systems accessible to broader populations in the future.

Compliance with ethical standards

Disclosure of conflict of interest

There is not conflict of interests.

Statement of ethical approval

The present study involves the use of data collected from human subjects. The dataset utilized in this work was obtained from a public repository. It is important to note that the dataset providers have already ensured that all necessary ethical considerations, permissions, and approvals were addressed during the data collection process. In this study, we did not conduct any data collection or associated activities ourselves. Instead, we relied on the publicly available dataset to perform our analysis and draw conclusions.

References

[1] Mindell, J.A.; Meltzer, L.J.; Carskadon, M.A.; Chervin, R.D. Developmental Aspects of Sleep Hygiene: Findings from the 2004 National Sleep Foundation Sleep in America Poll. Sleep Med. 2009, 10, 771–779.

[2] Hasan, M. J., Shon, D., Im, K., Choi, H. K., Yoo, D. S., & Kim, J. M. (2020). Sleep state classification using power spectral density and residual neural network with multichannel EEG signals. Applied Sciences, 10(21), 7639.

[3] Tarokh, L.; Carskadon, M.A. Developmental Changes in the Human Sleep EEG during Early Adolescence. Sleep 2010, 33, 801–809.

[4] Lucey, B.P.; Mcleland, J.S.; Toedebusch, C.D.; Boyd, J.; Morris, J.C.; Landsness, E.C.; Yamada, K.; Holtzman, D.M. Comparison of a Single-channel EEG Sleep Study to Polysomnography. J. Sleep Res. 2016, 25, 625–635.

[5] Chriskos, P.; Frantzidis, C.A.; Nday, C.M.; Gkivogkli, P.T.; Bamidis, P.D.; Kourtidou-Papadeli, C. A Review on Current Trends in Automatic Sleep Staging through Bio-Signal Recordings and Future Challenges. Sleep Med. Rev. 2021, 55, 101377.

[6] Mohebbi, M.; Ghassemian, H. Prediction of Paroxysmal Atrial Fibrillation Using Recurrence Plot-Based Features of the RR-Interval Signal. Physiol. Meas. 2011, 32, 1147.

[7] Senthilpari, C.; Yong, W.H. Epileptic EEG Signal Classifications Based on DT-CWT and SVM Classifier. J. Eng. Res. 2021, 10, N0 2A.

[8] IEEE Transmitter. Improving the Quality of Sleep with AI and Machine Learning. 21 May 2021. Available online: https://transmitter.ieee.org/improving-the-quality-of-sleep-with-ai-and-machine-learning (accessed on 10 October 2022).

[9] Imtiaz, S.A. A Systematic Review of Sensing Technologies for Wearable Sleep Staging. Sensors 2021, 21, 1562.

[10] Imtiaz, S.A.; Rodriguez-Villegas, E. Automatic Sleep Staging Using State Machine-Controlled Decision Trees. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) , Milan, Italy, 25–29 August 2015; IEEE: Manhattan, NY, USA, 2015; pp. 378–381.

[11] Sen, B.; Peker, M.; Cavusoglu, A.; Celebi, F.V. A Comparative Study on Classification of Sleep Stage Based on EEG Signals Using Feature Selection and Classification Algorithms. J. Med. Syst. 2014, 38, 18.

[12] Memar, P.; Faradji, F. A Novel Multi-Class EEG-Based Sleep Stage Classification System. IEEE Trans. Neural Syst. Rehabil. Eng. 2017, 26, 84–95.

[13] Santaji, S.; Desai, V. Analysis of EEG Signal to Classify Sleep Stages Using Machine Learning. Sleep Vigil. 2020, 4, 145–152.

[14] Bhusal, A.; Alsadoon, A.; Prasad, P.W.C.; Alsalami, N.; Rashid, T.A. Deep Learning for Sleep Stages Classification: Modified Rectified Linear Unit Activation Function and Modified Orthogonal Weight Initialisation. Multimed. Tools Appl. 2022, 81, 9855–9874.

[15] Tao, Y.; Yang, Y.; Yang, P.; Nan, F.; Zhang, Y.; Rao, Y.; Du, F. A Novel Feature Relearning Method for Automatic Sleep Staging Based on Single-Channel EEG. Complex Intell. Syst. 2022.

[16] Yulita, I.N.; Fanany, M.I.; Arymurthy, A.M. Sleep Stage Classification Using Convolutional Neural Networks and Bidirectional Long Short-Term Memory. In Proceedings of the 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS) , Bali, Indonesia, 5 June 2017; IEEE: Manhattan, NY, USA, 2017; pp. 303–308.

[17] Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.-K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circula*tion* 2000, 101, e215–e220.