Sleep Stage Classification Using CNN-LSTM and CNN-Transformer

Ruoyao Zhang

2025/12/02

Abstract

Automatic sleep stage classification plays a crucial role in analyzing sleep quality and diagnosing sleep disorders. This work investigates two modern deep learning architectures, CNN-LSTM and CNN-Transformer hybrid models, for classifying raw EEG signals into the five standard sleep stages: W, N1, N2, N3, REM. The Sleep-EDFx Sleep Cassette dataset is used for model training. A full preprocessing pipeline is developed, including EDF parsing, filtering, epoch extraction, and class imbalance mitigation via weighted cross-entropy. Both models are trained under identical five-fold cross-validation conditions to ensure unbiased comparison. We achieve mean macro-F1 scores of 0.743 (CNN-LSTM) and 0.741 (CNN-Transformer), comparable to or exceeding published baselines. Results show that Transformers do not significantly outperform LSTMs when restricted to single-epoch EEG input. These findings highlight the importance of incorporating temporal context for further model improvements.

1 Introduction

Sleep staging is the process of categorizing EEG-based polysomnographic (PSG) data into distinct physiological stages, each characterizing a sleep state. According to AASM standards [1], sleep is divided into five stages:

- Wake (W): categorized by high-frequency, low-amplitude EEG activity.
- N1: light transitional sleep with reduced alpha activity.
- N2: light sleep with sleep spindles and K-complexes.
- N3: deep slow-wave sleep dominated by delta activity.
- REM: mixed-frequency EEG with rapid eye movements and muscle atonia.

Manual sleep stage classification requires trained technicians and is both time-consuming and subjective [3]. Therefore, automated sleep stage classification has attracted strong interest among recent research [6].

Among the methodologies for sleep stage classification, deep learning approaches have shown promise. A widely used deep learning architecture is CNN layers followed by LSTM layers: CNNs can extract frequency-temporal features, while LSTMs capture sequence patterns. On the other hand, the Transformer model architecture is relatively novel to EEG signal classification, but frequently seen in modern deep learning models and acts as the foundation for large language models (LLMs) like ChatGPT, as transformers [5] excel at modeling long-range dependencies. This project investigates whether new approaches using Transformers outperform the classic LSTMs for single-epoch classification under matched experimental conditions.

2 Dataset

We use the Sleep Cassette (SC) subset of the Sleep-EDFx dataset [2]. The dataset includes EEG channels Fpz–Cz and Pz–Oz sampled at 100 Hz, along with hypnogram labels scored using the R&K manual [3]. After preprocessing (Section 4), we obtain 414,961 labeled 30-second epochs.

2.1 Class Distribution

Stage	**Count**
Wake (W)	285,433
N1	21,522
N2	69,132
N3	13,039
REM	25,835

Table 1: Final class distribution. Strong imbalance is present, especially in N1 and N3.

3 Preprocessing

Below are the steps we perform in the preprocessing pipeline:

1. EDF loading and annotation alignment.
2. EEG channel selection (Fpz–Cz and Pz–Oz).
3. Bandpass filtering (0.3–35 Hz), as recommended by AASM [1].
4. Epoching into 30-second segments.
5. Label merging and data cleaning.
6. Per-epoch normalization using:
$$x_{norm} = \frac{x - \mu}{\sigma + 10^{-6}}.$$

The final dataset shape obtained by this preprocessing pipeline is:

$$X \in \mathbb{R}^{414{,}961 \times 2 \times 3000}, \quad y \in \mathbb{R}^{414{,}961}.$$

4 Addressing Class Imbalance

To account for the high imbalance of the amount of data for different classes, we use weighted cross-entropy:

$$w_c = \frac{N}{n_c},$$

producing weights:

$$w = [0.2908,\ 3.8562,\ 1.2005,\ 6.3649,\ 3.2124],$$

This significantly improves minority class detection and prevents collapse into predicting mostly Wake.

5 Models

5.1 CNN-LSTM

Architecture:

Refer to Figure 1 for detailed diagram.

The CNN layers extract time–frequency features from raw EEG, and the bidirectional LSTM models temporal dependencies within each 30-second epoch. This architecture allows the model to capture both local spectral patterns and sequential dynamics, which are essential for distinguishing sleep stages such as N2 (spindles/K-complexes) and N3 (slow waves).

5.2 CNN-Transformer

Architecture:

Refer to Figure 2 for detailed diagram.

The CNN front-end mirrors that of the CNN–LSTM model to ensure a fair comparison. A Transformer encoder then applies multi-head self-attention to learn long-range temporal relationships within the extracted feature sequence. This model evaluates whether attention-based sequence modeling offers advantages over recurrent modeling for single-epoch sleep staging.

6 Training Procedure

Both models are trained with:

- Five-fold cross-validation (same folds for both models).
- Adam optimizer (lr = 1e-3).
- Weighted cross-entropy loss.
- Batch size: 64.
- 20 epochs per fold.

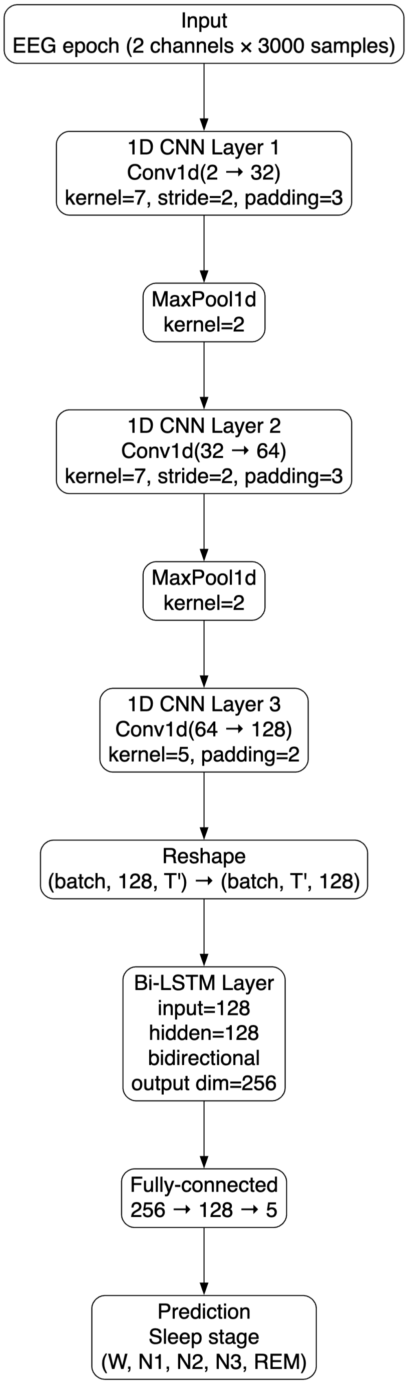

Figure 1: Detailed architecture of the CNN–LSTM model.

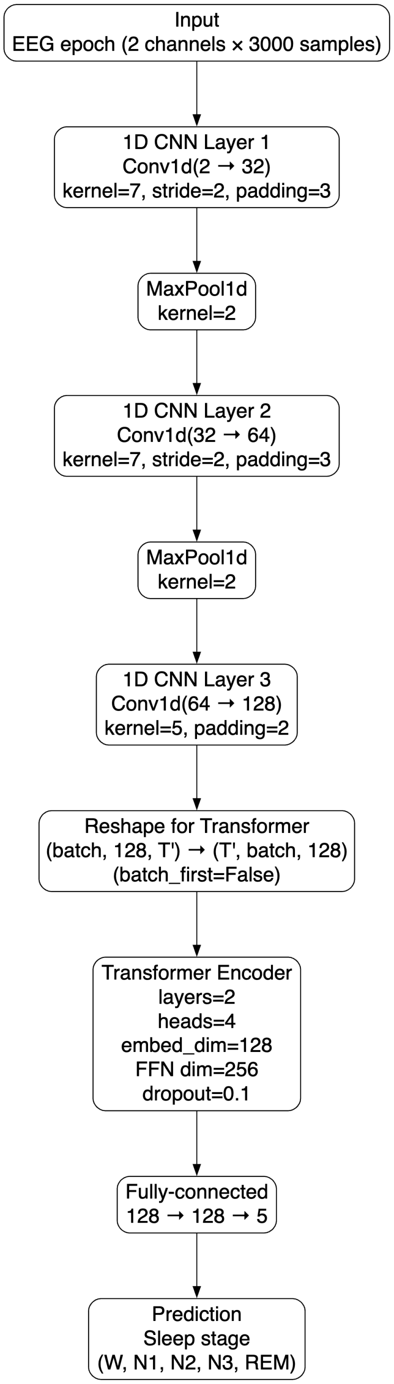

Figure 2: Detailed architecture of the CNN–Transformer model.

7 Results

7.1 CNN-LSTM

Fold	Acc	Macro F1
1	0.885	0.739
2	0.882	0.742
3	0.889	0.749
4	0.883	0.736
5	0.892	0.749
Mean	**0.886**	**0.743**

Table 2: Cross-validation results for the CNN–LSTM model.

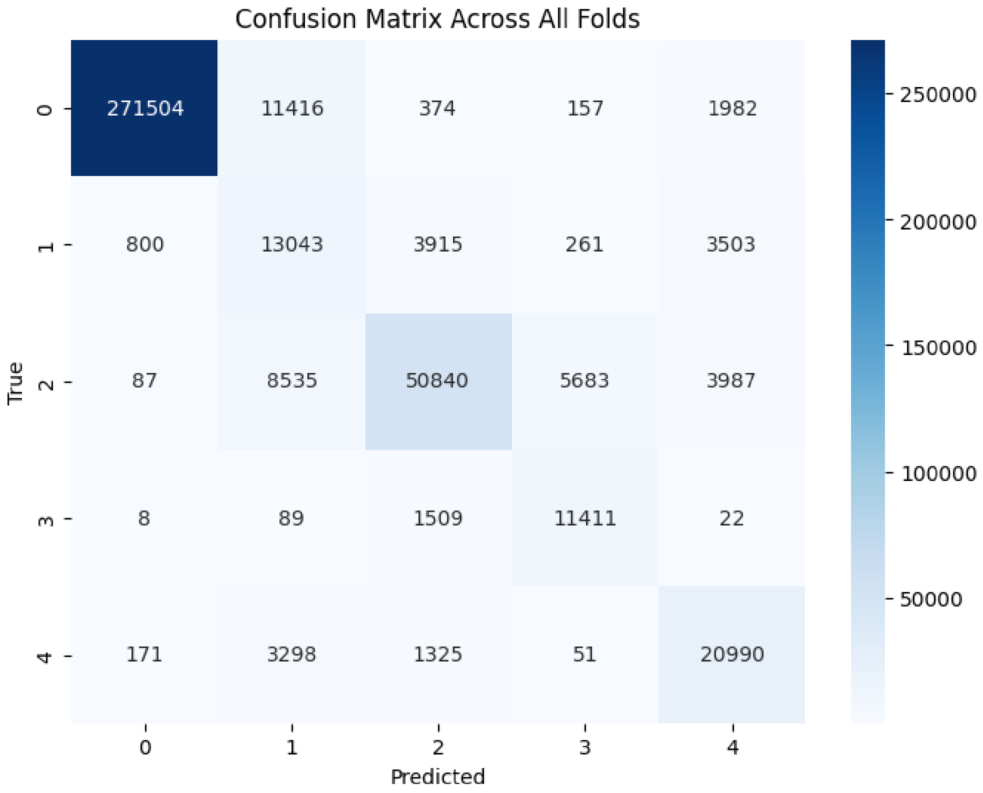

Figure 3: Confusion matrix for the CNN–LSTM model aggregated across all folds.

7.2 CNN-Transformer

Fold	Acc	Macro F1
1	0.895	0.747
2	0.892	0.754
3	0.872	0.727
4	0.874	0.740
5	0.881	0.737
Mean	**0.883**	**0.741**

Table 3: Cross-validation results for the CNN–Transformer model.

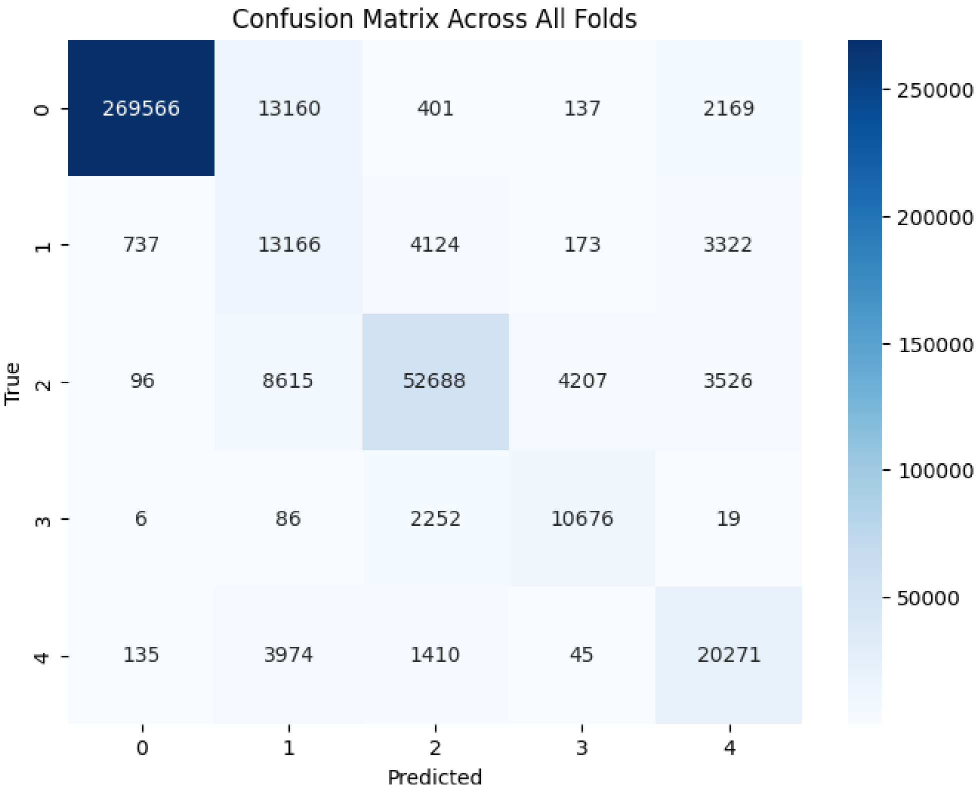

Figure 4: Confusion matrix for the CNN–Transformer model aggregated across all folds.

8 Discussion

Tables 2 and 3 show that both models achieve high overall accuracy of ≈0.88 and strong macro-F1 scores of ≈0.74. These results are competitive with established baselines such as DeepSleepNet [4]. Although LSTM and multi-head self-attention have large architectural differences, the two models produce nearly identical quantitative performance when they are provided single-epoch inputs.

The confusion matrices in Figures 3 and 4 provide deeper insight into class-wise behavior. Both models make excellent predictions on Wake, N2, N3, and REM stages, reflecting their ability to capture prominent EEG features such as alpha attenuation for Wake, spindles and K-complexes for N2, slow-wave activity for N3, and mixed-frequency REM patterns.

The most challenging stage for both models is N1. Since N1 is a transitional state with weak physiological markers and often seems similar to Wake or N2, models tend to struggle distinguishing these classes. Although the CNN–Transformer model shows slightly stronger performance on N3, and the CNN–LSTM performs slightly better on N1, the overall patterns of confusion are similar. This suggests that the primary limitations arise from the intrinsic difficulty of single-epoch classification rather than architectural differences between LSTMs and Transformers.

One possible explaination to why the Transformer architecture does not greatly outperform LSTM is that Transformers excel in modeling long-range temporal dependencies [5], but 30-second epochs of EEG data provide limited context for the self-attention mechanism to take advantage. Recent studies have shown that multi-epoch temporal context, as incorporated in SeqSleepNet and Transformer-XL variants, substantially improves N1 and REM performance. Thus, future work should investigate applying Transformers to sequences of multiple consecutive EEG epochs rather than isolated windows.

9 Conclusion

This work builds a complete preprocessing and classification pipeline for EEG-based sleep staging using CNN-LSTM and CNN-Transformer architectures. Both models achieve strong performance and show that architectural differences alone do not yield major improvements without incorporating additional temporal context.

10 Future Work

Future work could focus on comparing LSTM and Transformer-based architectures on multi-epoch temporal context windows.

References

[1] American Academy of Sleep Medicine, “The AASM Manual for the Scoring of Sleep and Associated Events,” 2007.

[2] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen and J. J. L. Oberye, ”Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG,” in IEEE Transactions on Biomedical Engineering, vol. 47, no. 9, pp. 1185-1194, Sept. 2000, doi: 10.1109/10.867928.

[3] A. Rechtschaffen and A. Kales, *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. 1968.

[4] A. Supratak, H. Dong, C. Wu and Y. Guo, ”DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG,” in IEEE Transactions on Neu-

ral Systems and Rehabilitation Engineering, vol. 25, no. 11, pp. 1998-2008, Nov. 2017, doi: 10.1109/TNSRE.2017.2721116.

[5] A. Vaswani et al., “Attention is All You Need,” *NeurIPS*, 2017.

[6] Liu, P., Qian, W., Zhang, H. et al. Automatic sleep stage classification using deep learning: signals, data representation, and neural networks. Artif Intell Rev 57, 301 (2024). https://doi.org/10.1007/s10462-024-10926-9