Application of Transformers for Sleep Stage Classification

David Vázquez-Lema, Diego Alvarez-Estevez, and Eduardo Mosqueira-Rey

CITIC, Universidade da Coruña, A Coruña, Spain
Correspondence: david.vazquez7@udc.es

DOI: *https://doi.org/10.17979/spu.23.c23*

Abstract: Transformer architectures have revolutionized the field of artificial intelligence. This work investigates the application of Transformers to the sleep stage classification problem. Despite promising results in recent studies, the clinical application of these methods remains limited. In this paper, we propose a novel model that replaces the LSTM component of a state-of-the-art CNN+LSTM architecture with a Transformer encoder, comparing its performance against our baseline and several state-of-the-art models. The results showed faster convergence, reduced complexity, and enhanced performance. Leveraging the Transformer architecture, we propose and investigate an interpretability method based on attention mechanisms. Finally, we evaluate the inter-database generalization performance of our model.

1 Introduction

Sleep is a vital biological function that impacts nearly every aspect of physical and mental well-being, but it's often overlooked. Sleep deprivation can lead to various health issues, from reduced cognitive performance to severe sleep disorders, highlighting the importance of accurate diagnosis and treatment.

The polysomnogram (PSG) is a key tool for diagnosing sleep disorders, monitoring signals such as brain waves (EEG), eye movements (EOG), and muscle activity (EMG). For sleep staging, experts analyze these signals in 30-second segments to classify them into five stages (W, N1, N2, N3, R), creating a hypnogram that visually represents the sleep pattern of a person. The analysis of the PSG is time-consuming and prone to human error, which can lead to scoring inconsistencies. To address this problem, automated methods have been developed to assist experts in sleep staging.

Early automatic methods relied on traditional deep learning architectures, such as Deep Neural Network (DNN), Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Since then, more complex architectures have emerged, some combining traditional architectures, such as CNN+LSTM, while others introducing entirely new designs, such as transformer architectures. For a more detailed description of the state-of-the-art, we suggest revising the paper of Fiorillo et al. (2019).

This research explores the use of deep learning, specifically a novel CNN+ Transformer architecture, to improve the results achieved by state-of-the-art models. We also aim to enhance inter-database generalization and add interpretability capabilities by leveraging the Transformer's attention mechanisms to explain the model's decisions.

2 Materials and methods

Our proposed approach was inspired by the research done by Anido-Alonso and Alvarez-Estevez (2023). They proposed their own innovative pipeline for sleep stage classification,

introducing the novel aspect of inter-database validation across several well-known databases. During our work, we have used their proposed CNN+LSTM architecture as our baseline model. We have also modified it to integrate the innovative transformer into its architecture.

The schema shown in Figure 1 describes the blocks that make up our proposed architecture. The first two blocks are identical to those described in the baseline architecture. The main innovation is the replacement of the LSTM block for a transformer block.

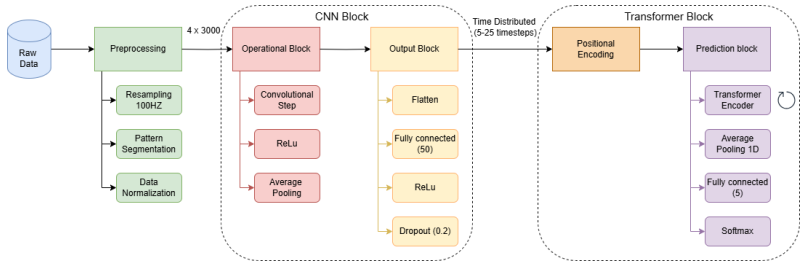

Figure 1: Overview of the CNN+Transformer architecture.

The first component of this block is the positional encoding, responsible for injecting spatial information into the input vectors. These vectors are then fed into a transformer encoder, which can be repeated multiple times to extract more complex relationships. The main advantage compared to an LSTM is its capacity to handle longer sequences, thanks to its attention mechanism. The stacked encoders are followed by a global average pooling 1D layer, responsible for dimensionality reduction. Finally, a fully-connected layer followed by a softmax activation function will produce the final output.

2.1 Interpretability

This work presents a method for interpreting transformer models by analyzing attention values. The method focuses on extracting attention scores from the final multi-head attention layer to understand how the model makes its decisions. The proposed method works by combining the attention matrices of different heads into a single matrix. This unified matrix helps identifying the most influential parts of the input sequence. In this work we experiment with two configurations to split the network's input signal into different attention sequence lengths. Experiments are described in more detail in the following section.

3 Experiments

This section details the different experiments conducted during this work. A total of three architectures were compared: the CNN+LSTM baseline, the proposed CNN+ Transformer (CNN+TF) without positional encoding, and the complete CNN+ Transformer architecture (CNN+TF_PE). For each architecture, we tested configurations that combined different split values (5 and 25). Table 1 summarizes the resulting models.

Split values divide the input vector of dimensions 4×15000 to generate the input sequences. If we select a value of 5, we will obtain 5 sequences of dimension 4×3000, each sequence containing 30s of data. In contrast, a split value of 25 results in 25 sequences of dimensions 4×600, each sequence containing 6s of data. These split values were selected for specific reasons: a value of 5 was found to be the most suitable in the study conducted by Alvarez-Estevez and Rijsman (2021), while a split of 25 was chosen to evaluate the ability of Transformers to handle longer sequences, as the use of Transformers in a short-length split sequence of 5 may be unnecessary and computationally inefficient. Split values are also related to interpretability, influencing how the importance of each input element is determined:

1. **Split of 5**. Each input element corresponds to one 30s epoch. Averaging each attention matrix results in a 5×5 matrix, and the importance of each element is extracted from the third row, which corresponds to the central epoch.
2. **Split of 25**. Each 30s epoch is divided into 5 parts. Averaging each attention matrix results in a 25×25 matrix. Since the target is the central epoch (5 elements), their rows are averaged to obtain a 25-length importance vector that represents each element's contribution to the prediction.

Other parameters like the use of positional encoding, the feedforward network dimension, and the number of times the transformer structure is repeated, are also stated in the table. A common batch size of 100 was selected because it provided a good balance between training stability and computational efficiency, allowing the models to converge effectively without exceeding memory limitations. This memory problem may not occur with LSTM architectures, but can arise with Transformers due to their higher memory consumption. In order to avoid overfitting, an early stopping value of 10 was selected in all experiments. In the case of models using a LSTM, we set the number of LSTM units at 1000, as we found during experimentation that using only 100 units as in the original study Alvarez-Estevez and Rijsman (2021) was somewhat limiting. After some experimentation with the transformer models, the head size, the number of heads and the dropout values were fixed to the values 10, 5 and 0.2, respectively.

Table 1: Summary of model configurations for the experiments

	P.Enc.	FF	Blocks	Split
CNN+LSTM_5	-	-	-	5
CNN+LSTM_25	-	-	-	25
CNN+TF_5	NO	1024	1	5
CNN+TF_25	NO	1024	1	25
CNN+TF_PE_5	YES	524	4	5
CNN+TF_PE_25	YES	1024	1	25

The resulting models were evaluated using two sleep staging databases: Sleep Heart Health Study (SHHS) and the PolySomnoGraphic Inter-scorer Performance Assessment database (PSG-IPA). More specifically, SHHS was used to train and evaluate the model, while the whole PSG-IPA was used to assess the inter-database generalization performance. Regarding the partitioning of the SHHS database, 20% of the total data is used for testing, while the remaining 80% is further divided into training and validation sets, comprising 80% and 20%, respectively.

4 Results

Table 2 presents the results of the experiments performed in this work, including the evaluation metrics, the number of model parameters, and the number of training epochs required. For evaluating the models, we used Cohen's Kappa index, a metric considered more robust than traditional alternatives, such as accuracy, because it accounts for the possibility of agreement occurring by chance Cohen (1960). This allows for a better performance comparison in imbalanced problems such as sleep stage classification. Moreover, Cohen's Kappa is the standard metric used in the literature to assess inter-rater agreement in sleep scoring tasks. We present the results obtained by evaluating the model in both the SHHS database testing partition and the entire PSG-IPA database, shown in the table under the columns *Test Kappa* and *External Kappa*, respectively. The number of parameters of the model is also stated in the table, which shows the different complexities of the models. Finally, we report the early stopping epoch at which the model finished the training process.

Table 2: Summary of experimental results

	Test Kappa	External Kappa	Number of Parameters	Epochs
CNN+LSTM_5	0.824	0.545	6.673.911	62
CNN+LSTM_25	0.843	0.536	4.753.911	83
CNN+TF_5	0.788	0.461	2.585.953	10
CNN+TF_25	0.807	0.477	665.953	16
CNN+TF_PE_5	0.815	0.475	2.725.575	19
CNN+TF_PE_25	0.853	0.548	665.953	31

Table 3 shows the comparison of the best performing model developed during this work with other well-known state-of-the-art models. For a more realistic comparison, the selected models were trained in the same database, although some were trained on SHHS-1 instead of SHHS-2 and may not have used the same input channels. Additionally, architectural diversity was considered by including models trained with CNN+LSTM as well as others with transformers. All values were taken directly from their respective reports.

Table 3: Performance comparison with other models trained on SHHS database and validated with inter-database generalization on PSG-IPA database.

Model	Test Kappa	External Kappa
CNN+TF_PE_25	0.853	0.548
Alvarez-Estevez and Rijsman (2021)	0.820	-
Anido-Alonso and Alvarez-Estevez (2025)	0.790	0.480
Phan et al. (2022)	0.828	-
Pradeepkumar et al. (2024)	0.792	-

4.1 Interpretability

In this section, we interpret the results achieved by the best performing model, CNN+ TF_PE_25, by applying the proposed method described in Section 2.1. Figure 2 illustrates the importance of an input sequence labeled W. The graphs shown in this figure present the input channels and the importance of each element of the input sequence. The most significant elements in the shown example are those corresponding to the central epoch (10-15) or those in close proximity. This outcome is logical, as the region that we aim to predict is the center one, so the closest elements should have a greater impact on the final decision. If we further analyze the input channels, we can observe that W stages tend to present movement activity in the EOG, and relatively high amplitude in the EMG, indicating continuous eye movements and muscle activation. This activity is characteristic of periods of wakefulness.

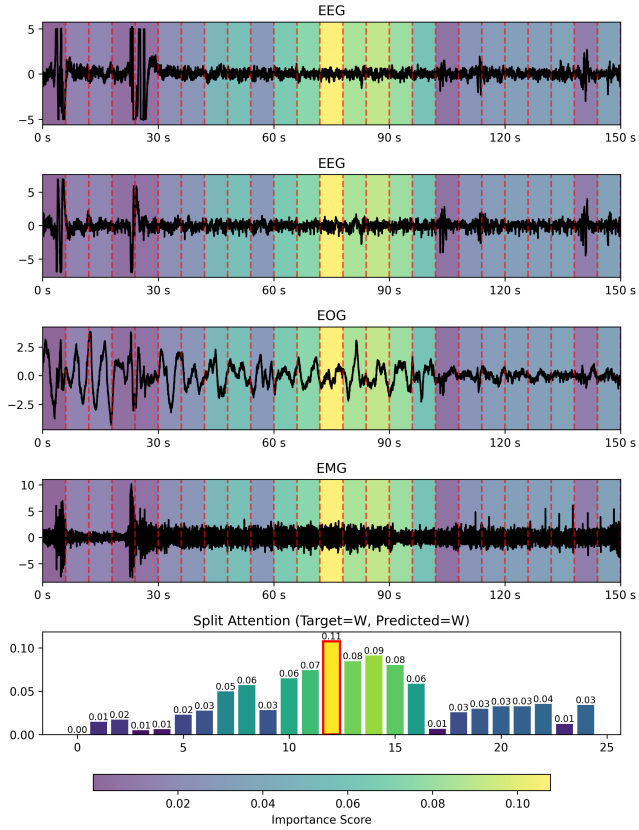

Figure 2: Interpreting results for an input sequence with a split value of 25 for a W stage. The first four plots represent the input signals for two EEG, EOG, and EMG channels. The bar plot shows the importance of each element in the input sequence. The color code provides a visual representation of the importance of each element across the different graphs, where yellow tones indicates more importance and purple tones less importance to the final classification.

5 Discussion

This work explored several deep learning architectures for automatic sleep stage scoring in polysomnographic recordings. In particular, we evaluated and modified an existing state-of-the-art architecture to incorporate the novel and powerful transformer. To this end, we have swapped one of the main model blocks, the LSTM one, with a transformer. To better understand these changes, the number of input sequences was increased by decreasing the length of each sequence, trying to evaluate the capacity of the transformers to handle long sequences.

Regarding the variation in length in the input sequence, the results presented in Table 2 manifested a significant improvement in sleep stage classification, particularly with the transformer architectures. The increase in sequence length benefits Transformers by providing richer contextual information and capturing better long-range dependencies, resulting in improved generalization capabilities. Remarkably, the Transformer architecture with positional encoding significantly increased the performance on both Test Cohen's Kappa and External Cohen's Kappa, increasing from 0.815 to 0.853 and from 0.475 to 0.548, respectively. In contrast, the limitations of the CNN+LSTM architectures are evidenced by this variation in the sequence length. Although slightly improving Test Cohen's Kappa score, from 0.824 to 0.843, it decreased the External Cohen's Kappa score, from 0.545 to 0.536. Although the variation in performance is not as pronounced as in the Transformer models, the slight decrease in inter-database generalization capabilities highlights the inherent limitations of the LSTM.

With regard to the performance of the models, noticeable variation can be seen between the models that use positional encoding and those that do not. Despite similar performance in the External Cohen's Kappa metric across transformer models, with the exception of CNN+TF_PE_25, those that incorporated positional encoding exhibited significantly improved performance in the Test Cohen's Kappa metric. These results confirm that the order of the sequences is really important for time-series classification problems, particularly for sleep stage classification.

A significant drop in performance was found when evaluating the inter-database generalization capabilities of our models. Despite the existence of standard clinical guidelines, inter-database variability can still arise due to differences in recording and scoring methods, population demographics, or simply human interpretation, leading to inconsistent results and challenges in model generalization across datasets. Although this topic rarely appears in the literature, some studies have proposed some ways to mitigate its effects. The simplest approach to this problem would be to create a large, heterogeneous database by combining multiple common databases. However, other challenges can arise from the need to manage such a large database, including high costs, data privacy concerns, and scalability problems. Alternatively, the use of federated learning Anido-Alonso and Alvarez-Estevez (2023) and ensemble methods Alvarez-Estevez and Rijsman (2021) was proposed to mitigate the previously mentioned problems, taking advantage of their great scalability and flexibility, showing promising results.

The number of trainable parameters might also influence the achieved results. The baseline CNN +LSTM models have a higher number of parameters, theoretically giving them greater capacity to learn more complex patterns from the data, but with the drawback of being more susceptible to overfitting. In contrast, the tested Transformer models tend to have considerably fewer trainable parameters. The reduction in parameters, in conjunction with their great parallelism capabilities, results in a significantly faster training.

When considering the Test Kappa performance on the SHHS database, our proposed architecture achieves slightly better results than the state-of-the-art models that we have considered. This improvement in performance can be attributed to multiple factors: (i) The use of a larger sequence length compared to the other models, which were usually trained with 30s splits; (ii) The addition of transformers to the architecture that, in relation to the previous factor, can create better representations; Finally, (iii) the addition of channels, such as EOG and EMG, significantly enhances the ability to identify certain sleep stages, such as REM or W. A significant improvement was also observed with respect to the External Kappa score. However, the limited

literature on this topic makes it difficult to find other studies in similar settings. Compared to the only study that performed inter-database generalization with PSG-IPA Anido-Alonso and Alvarez-Estevez (2025), our proposed model clearly outperforms the reported results. Other factors, such as data partition, might have slightly influenced the results obtained.

In terms of interpretability, the proposed method consolidates the multiple attention matrices from each head into a single, unified representation. This consolidation offers insight into the most influential elements in the decision-making process to produce a certain output. In terms of the split value, longer values offer increased resolution, allowing for a more precise identification of the most relevant elements within the input signal. In contrast, smaller values limit the importance to longer epochs, restricting the interpretability capabilities of the model.

It is also important to discuss some potential limitations of our work. In this regard, one possible issue with the use of Transformers is their computational demands and substantial memory requirements for storing all intermediate operations prior to returning a result. Because of this added cost, in this work we were unable to perform multiple repetitions of our experiments to provide statistical significance to support our results. Regardless, our data suggests that the proposed architecture can achieve results at least similar to state-of-the-art models while being less complex and training faster. Regarding interpretability, and despite considerable research efforts to convert Transformer models into white boxes, discussion remains open. The presented approach is one step forward, but decision-making remains difficult due to the complexity of the underlying problem. Future work will be carried out to address these issues.

6 Conclusions

Our work shows how our proposed CNN+Transformer approach outperforms the baseline architecture and other state-of-the-art models in terms of Test Cohen's Kappa score, while also obtaining similar results in terms of inter-database generalization. In addition, our proposed approach offers other advantages, such as reduced training time and model complexity. Particularly, the use of larger splits tends to enhance the performance of transformer models. In addition to the proposed architecture, we attempt to eliminate the black-box problem that most deep learning models have by proposing and testing an interpretability method that makes use of the attention mechanisms.

Acknowledgments

This work has been supported by project PID2023-147422OB-I00, funded by MCIU/AEI/10.13039/501100011033 and by the European Regional Development Fund (ERDF) program, and by the Xunta de Galicia (Grant ED431C 2022/44), supported by ERDF. CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the ERDF Galicia 2021-27 operational program (Ref. ED431G 2023/01). SMR has received funding from Xunta de Galicia (grant ED481A 2023/008). DAE has also received support from project RYC2022-038121-I, funded by MCIN/AEI/10.13039/501100011033 and European Social Fund Plus (ESF+), and project ED431F 2025/35 from Xunta de Galicia.

Bibliography

D. Alvarez-Estevez and R. M. Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PLOS One*, 16(8):e0256111, 2021.

A. Anido-Alonso and D. Alvarez-Estevez. Decentralized data-privacy preserving deep-learning approaches for enhancing inter-database generalization in automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 27(11):5610–5621, 2023.

A. Anido-Alonso and D. Alvarez-Estevez. Multi-task deep-learning for sleep event detection and stage classification. In *2025 IEEE Symposium on Computational Intelligence in Health and Medicine Companion (CIHM Companion)*, pages 1–5, 2025.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

L. Fiorillo, A. Puiatti, M. Papandrea, P.-L. Ratti, P. Favaro, C. Roth, P. Bargiotas, C. L. Bassetti, and F. D. Faraci. Automated sleep scoring: A review of the latest approaches. *Sleep Medicine Reviews*, 48:101204, 2019. URL *https://www.sciencedirect.com/science/article/pii/S1087079218301746*.

H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467, 2022.

J. Pradeepkumar, M. Anandakumar, V. Kugathasan, D. Suntharalingham, S. L. Kappel, A. C. De Silva, and C. U. Edussooriya. Towards interpretable sleep stage classification using cross-modal transformers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024.