Comparative Study of Traditional Machine Learning and Deep Learning with LLMs for Sleep Staging with Time Series

Rushil Baindla[1]

Received September 17, 2025
Accepted August 22, 2026
Electronic access September 30, 2026

Automated sleep staging from polysomnographic signals is essential for diagnosing sleep disorders, yet the potential of large language models (LLMs) for this task remains underexplored. This study benchmarks traditional deep learning architectures (RNN, LSTM, GRU, Transformer) against zero-shot LLM approaches (GPT for Time Series, GPT-Neo) for EMG-based sleep stage classification using a public dataset of 100 subjects, yielding 26,992 windowed samples evaluated via 3-fold cross-validation. The GRU achieved the highest accuracy (75.96%), substantially outperforming competing models (73.9-74.62%). Zero-shot LLMs performed near chance level, with GPT-Neo showing an AUC of 0.50. These findings indicate that task-specific deep learning models remain preferable for biosignal classification, while zero-shot LLMs require domain adaptation before clinical deployment.

Introduction

Sleep disorders such as insomnia and obstructive sleep apnoea are prevalent and clinically significant, yet they still lack definitive cures. Advances in ML have enabled automated sleep staging, classifying epochs into Wake, N1, N2, N3, and REM, using EEG, ECG, EMG, and EOG[1–3]. Accurate staging is foundational for downstream detection of irregularities and prediction of sleep-related complications[4].

Here, the study focuses on EMG-based sleep staging using a public dataset (see the Data section). The study compares deep learning architectures with zero-shot LLM baselines to evaluate performance. Based on prior evidence that zero-shot LLMs underperform specialised models on biosignals[2,5,6] the study hypothesised that trained models would outperform zero-shot approaches on time-series classification. The study selected four deep learning architectures that represent distinct approaches to sequential modelling. Standard RNNs serve as a foundational baseline, capturing temporal dependencies via recurrent connections, but have difficulties from vanishing gradients on long sequences. LSTMs address this limitation through gating mechanisms that regulate information flow, making them a standard choice for biosignal classification tasks, including sleep staging and seizure detection[7]. GRUs offer a computationally lighter alternative with fewer parameters while maintaining comparable performance on many sequence tasks. Transformers utilize self-attention to model long-range dependencies without recurrence, to produce results across varying sequence domains[8]. These architectures enable systematic comparison of architectural complexity against classification performance on physiological time series. The study has three key contributions: (1) a reproducible ML pipeline for multi-channel time series; (2) benchmarks contrasting traditional models and zero-shot LLMs for sleep staging; and (3) insights into the conditions under which recurrent architectures outperform attention-based models on short physiological time series[9,10].

Related Work

The literature review section covers the subtopics explored in the data. The subtopics being heart activity (ECG), brain activity (EEG), and overall innovation. First, despite LLM-based approaches being promising for ECG and EEG interpretation, the most successful applications of LLMs rely on task specific fine tuning rather than just zero shot inference. Second, EMG is still under explored when compared to EEG and ECG, specifically for differentiating REM sleep from awakeness. Third, there are few direct comparisons between zero shot LLMs and existing sequential architectures on the same biosignal datasets, which limits the evaluation of practical advantages of LLMs over existing methods. The study fills these gaps by benchmarking zero shot LLMs against a systematic progression of recurrent and attention based models on EMG-based sleep staging, and providing empirical evidence on the current viability of unadapted LLMs for physiological time-series classification.

A study compared both machine and deep learning on polysomnographic signals and ultimately found that deep learning outperformed traditional classifiers across numerous sleep staging benchmarks[11]. Convolutional Neural Network-

[1] *Emerald High, USA*

based sleep staging on EEG signals from the Sleep EDF dataset received strong performance scores for classification, indicating a competitive baseline value for physiological data[12]. Similarly, end to end networks that combine both CNN and RNN for sleep data analysis showed improvements in accuracy[13]. Traditional machine learning analysis of solely EMG features from the Sleep EDF database achieved over eighty percent accuracy, indicating that EMG alone is a sufficient signal for classifying sleep stages even when underused by contemporary literature[14]. Combinations of EEG, EOG, and EMG signals, as seen in multimodal machines, demonstrate the additional value of multiple data channels for overall classification[15]. GRU based sleep staging using EEG channels has been used for its computational efficiency as a recurrent model. These models can match or even surpass more complicated architectures with specific feature selection[16].

Before the widespread usage of deep learning, sleep staging was based on a set of handcrafted features extracted from polysomnographic signals. These signals include spectral power, Hjorth parameters, and entropy measures, combined with classifiers such as support vector machines and random forest algorithms. Convolutional neural networks managed to reach strong performance by learning hierarchical features directly from raw signals with architectures like DeepSleepNet and U-Time, setting the benchmarks for public datasets[17,18]. A recent review of automated sleep staging approaches situates these architectures within the broader evolution of the field[19]. These techniques are still competitive and are still commonly used in clinical software. They provide the baseline against which newer architectures should be measured. The study's choice of RNN, LSTM, GRU and Transformer models continues this trajectory towards attention-based sequential modelling. Additionally, the study includes zero shot LLMs to see if general purpose language models can outperform task specific alternative models.

Heart Activity

LLM-based and foundation-model approaches for ECG are emerging. ECG-LM demonstrates that LLMs conditioned on ECG representations can support clinically meaningful interpretation tasks[20]. Zero-shot classification with multimodal prompts remains challenging but can be improved with test-time clinical knowledge[21]. A recent review synthesises ECG foundation-model progress, highlighting advantages in cross-task transfer and persistent gaps in data diversity and evaluation[22]. These advances inform the study's expectations about LLMs' strengths/limitations on physiological time-series.

Brain Activity

LLMs for EEG classification and interpretation are promising but not yet consistently superior to task-specific models. EEG-GPT explored LLM capabilities for EEG labelling/explanation, finding benefits mainly with adaptation and structure imposed on inputs[5]. For motor-imagery EEG, GPT-4-o underperformed traditional models by ~ 0.15 on accuracy and AUC, underscoring zero-shot limitations[6]. Comparative evaluation of both LSTM and GRU on raw EEG data found that both models were capable of predicting cognitive states, supporting their inclusion as baselines in signal classification[23]. More broadly, surveys outline four application areas foundation models, language decoding, media generation, and data management, and emphasise the need for richer EEG datasets and rigorous evaluation[2]. Clinically, EEG remains valuable for monitoring and prognosis in traumatic brain injury, motivating reliable staging/abnormality detection pipelines[4]. Recent work also targets personalised health summaries from EEG-image multimodal data[24] and thought-to-text pipelines that align EEG encoders with LLMs[25]. In sleep/attention contexts, combining behavioural and EEG signals with LLMs is a growing direction[1].

Advancements in AI

Time-series models that explicitly handle cross-channel dependencies and noise can outperform generic baselines. Temporal dynamic graph networks such as TodyNet learn inter-variable relations and have shown strong gains on multivariate benchmarks[7]. Cross-interaction refinement (CrossGNN) tackles noisy multivariate sequences by modelling cross-signal structure[10]. In clinical time-series, multi-resolution spatiotemporal graph learning can improve classification under distribution shifts[9]. Beyond physiology, multimodal transformer designs for human motion prediction illustrate how architectural choices and pretraining influence downstream performance[8]. Together, these findings support the paper's choice of strong task-specific baselines and caution against assuming zero-shot LLMs will excel without adaptation[3]. Literature today evaluating zero-shot time-series LLMs and broader LLMs applied to time series consistently finds that limitations in structural reasoning and gaps compared with task-specific models remain without fine-tuning[26,27]. A comprehensive review of transformer architecture across multiple biosignals established attention based models as a viable option for physiological time series analysis, even with challenges in persistent generalization[28].

Data

The study uses the Sleep-EDF Expanded (Sleep-EDFx) Cassette dataset[29], publicly available through PhysioNet under the Open Data Commons Attribution License. The dataset comprises polysomnographic recordings from 153 subjects (305 approximately 20-hour sessions), sampled at 100 Hz.

Each recording includes EEG (Fpz-Cz, Pz-Oz), EOG, and chin EMG channels, with expert-annotated hypnograms labelling 30-second epochs as Wake (W), N1, N2, N3, or REM.

Label distribution was recorded as Wake = 69% (18,617), Sleep = 31% (8,375). The dataset exhibits substantial class imbalance. As consistent with other analyses of the Sleep EDFx cassette, the data are mainly composed of N2 with a smaller proportion of N1.

Raw EDF files were segmented into 30-second epochs aligned with hypnogram annotations. EMG channels were extracted. A 0.5–40 Hz bandpass filter was applied to the EMG channel before feature extraction. This raw EMG signal was used directly. Epochs containing artefacts or unlabelled segments were excluded, yielding 26,992 windowed samples from 100 subjects. The analysis was conducted on 100 of the 153 available subjects because of GPU memory and runtime constraints inherent to the cloud-based training environment. All of the available recordings within this subset were included in full.

The data was partitioned using 3-fold cross-validation with subject-level splits done through GroupKFold. This process ensures no individual subject's data appears in both training and test folds within any fold. Across the three folds, 67 subjects were used for training and 33 for testing, corresponding to each fold. 5 minutes of sleep per sample was represented as 30-second epochs of 10 non-overlapping epochs. The label was assigned to the final epoch in each window.

Methods

To address the research question, a combination of quantitative analysis with programming-based benchmarking was required. The models were evaluated on both accuracy and training speed, with comparisons made relative to one another, as the overall performance was expected to be modest given the complexity of the task.

The study compares recurrent neural networks (RNN), long short-term memory (LSTM), gated recurrent units (GRU), and a transformer against zero-shot LLM baselines (GPT for Time Series, GPT-Neo). Recurrent models are well-suited to shorter sequences and can train efficiently; LSTMs add gating to capture longer dependencies; GRUs provide a compact alternative; transformers can capture long-range dependencies but typically require more data and compute [9,10]. LLMs were evaluated zero-shot without tuning, reflecting common practical constraints in healthcare [3,5,6].

All the models were trained for 20 epochs using the Adam optimizer on a batch size of 32. All models stopped improving before the limit. The study utilized 3 fold cross validation as stated above in the Data section. All scaling and gap filling were done using only the training portion to avoid the test results being influenced. For each 30-second epoch, the program recorded a five-number feature vector representing signal power, which were grouped into windows of 10 consecutive epochs.

Hyperparameter	RNN	LSTM	GRU	Transformer
Layer 1 units	64	64	64	embed_dim=64
Layer 2 units	32	32	32	ff_dim=64
Attention heads	—	—	—	2
Dropout	0.2	0.2	0.2	0.2
Dense units	16	16	16	16
Optimizer	Adam	Adam	Adam	Adam
Batch size	32	32	32	32
Epochs	20	20	20	20
Input shape	(10, 5)	(10, 5)	(10, 5)	(10, 5)

Below is a visual representation of an LSTM unit. LSTM functions through 3 main gates, and both long-term memory cell states and a short-term immediate output within its hidden state.

All deep learning models received input sequences of length 10 with 5 features per timestep corresponding to the five PSD frequency bands.

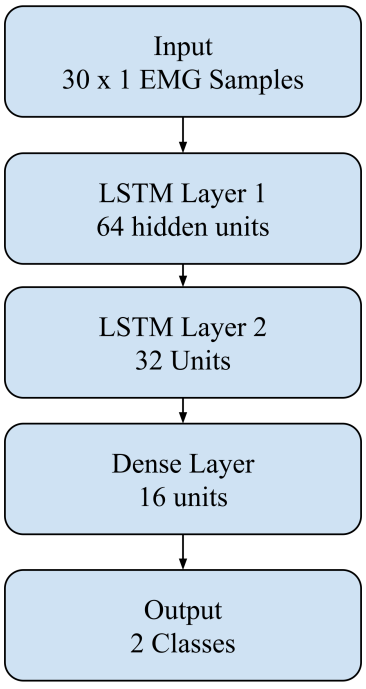

Figure. 1 LSTM Model for sleep stage classification.

In addition to these deep learning models, zero-shot large language models (LLMs) were tested, specifically GPT for Time Series and GPT-Neo. To apply GPT based models to EMG signals, raw time-series values were normalized to zero

mean and then converted to separate strings. Each input sequence was truncated or chunked to fit within the model's context window (1024 tokens for GPT-2, 2048 tokens for GPT-Neo). Prompts followed the structure: "an instruction specifying the task and candidate labels, followed by the raw signal values, followed by a classification query, where the instruction specified the task and candidate labels". The model's outputs were parsed by extracting the predicted class label from the generated text. Responses not matching a valid label were treated as excluded from any further evaluation. These models were provided with preprocessed data arrays, and their outputs were compared directly with the results of the neural network frameworks. The comparison enabled an assessment of whether LLM-based approaches could offer competitive performance for EMG sleep stage classification.

Zero-shot evaluation of GPT for Time Series and GPT-Neo required converting continuous EMG signals into text-compatible formats. Preprocessed EMG segments (length: 30 samples) were first normalised to zero mean and unit variance. Numerical values were then converted to string representations with 3 decimal places, separated by commas.

Each input was framed using the following prompt template:

[SYSTEM/INSTRUCTION]

You are a sleep staging classifier. Given the following EMG signal values, classify the sleep stage as either 0 or 1.

[INPUT]

EMG values: [v1, v2, v3, ..., vn]

[QUERY]

Sleep stage:

Models were evaluated in a strictly zero-shot setting with no in-context examples. Generation temperature was set to 0.7 for GPT-Neo with majority vote across 5 return sequences. GPT-2 used greedy decoding. The responses were parsed by extracting the first occurrence of a valid class label. Any output not containing a recognised label was excluded from evaluation. For sequences larger than the model's context window (2048 tokens for GPT-Neo), signals were truncated to the most recent tokens' worth of signal values. This truncation may limit the model's access to relevant temporal context and represents a known constraint of applying fixed context LLMs to extended time-series data.

Results

Results are reported as an average across all three cross validation folds. Classification reports are provided for each deep learning architecture and LLM baseline.

LSTM — 3-Fold CV Classification Report

Time Series Visualization for all modes relative to their prediction values. The following figures, Figures 3-6, demonstrate predictions across the sequence of EMG recordings.

Table 1 LSTM — 3-Fold CV Classification Report

	Precision	Recall	F1-Score	Support
0 (Wake)	0.7657	0.9105	0.8318	18617
1 (Sleep)	0.6566	0.3805	0.4818	8375
Accuracy			0.7460	26992
Macro Avg	0.7111	0.6455	0.6568	26992
Weighted Avg	0.7318	0.7460	0.7232	26992

Confusion Matrix: [[16950, 1667], [5188, 3187]]
AUC (macro): 0.7302 APRC (macro): 0.5860

Table 2 RNN — 3-Fold CV Classification Report

	Precision	Recall	F1-Score	Support
0 (Wake)	0.7607	0.9221	0.8337	18617
1 (Sleep)	0.6722	0.3551	0.4647	8375
Accuracy			0.7462	26992
Macro Avg	0.7165	0.6386	0.6492	26992
Weighted Avg	0.7332	0.7462	0.7192	26992

Confusion Matrix: [[17167, 1450], [5401, 2974]]
AUC (macro): 0.7255 APRC (macro): 0.5843

Table 3 GRU — 3-Fold CV Classification Report

	Precision	Recall	F1-Score	Support
0 (Wake)	0.7719	0.9247	0.8414	18617
1 (Sleep)	0.7011	0.3927	0.5034	8375
Accuracy			0.7596	26992
Macro Avg	0.7365	0.6587	0.6724	26992
Weighted Avg	0.7500	0.7596	0.7366	26992

Confusion Matrix: [[17215, 1402], [5086, 3289]]
AUC (macro): 0.7404 APRC (macro): 0.6168

Table 4 Transformer — 3-Fold CV Classification Report

	Precision	Recall	F1-Score	Support
0 (Wake)	0.7497	0.9332	0.8314	18617
1 (Sleep)	0.6743	0.3075	0.4223	8375
Accuracy			0.7390	26992
Macro Avg	0.7120	0.6203	0.6269	26992
Weighted Avg	0.7263	0.7390	0.7045	26992

Confusion Matrix: [[17373, 1244], [5800, 2575]]
AUC (macro): 0.7195 APRC (macro): 0.5812

Each plot compares predicted sleep probabilities to ground truth labels across 100 consecutive windows. In addition to the classification metrics done in the classification reports, these graphs illustrate prediction across different time series segments.

Below are the ROC curves for all traditional models. The x-axis represents the false positive rate and the y-axis the true positive rate. The area under the curve (AUC) represents overall model discrimination ability, with 0.5 indicating random

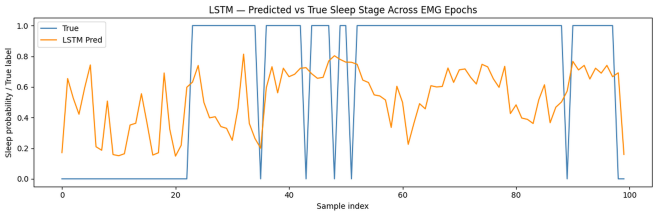

Figure. 2 LSTM Predicted vs Ground Truth Sleep Stage Across EMG Epochs

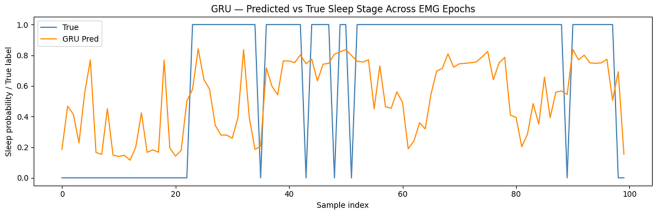

Figure. 3 GRU Predicted vs Ground Truth Sleep Stage Across EMG Epochs

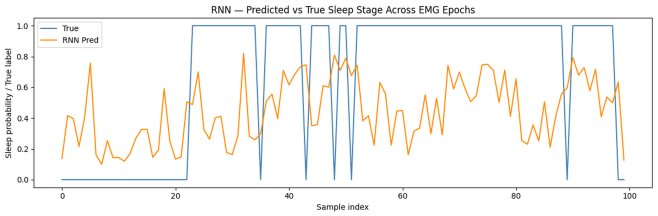

Figure. 4 RNN Predicted vs Ground Truth Sleep Stage Across EMG Epochs

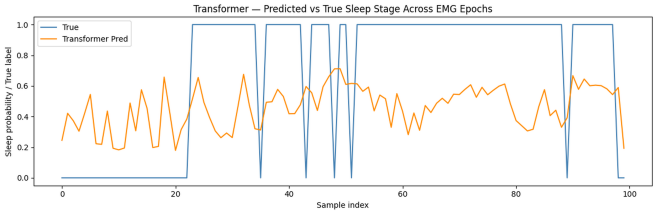

Figure. 5 Transformer Predicted vs Ground Truth Sleep Stage Across EMG Epochs

chance-like performance and 1.0 indicating a perfect classification. All four models substantially exceed the 0.5 baseline, with GRU achieving the highest AUC of 0.7404. Lastly, the Area under the Curve for an ROC graph represents the ability of the model, with an area of 0.5 being essentially guessing between binary trues and falses.

Table 5 GPT-2 (Time Series) — Zero-Shot Classification Report

	Precision	Recall	F1-Score	Support
0 (Wake)	0.5556	0.8333	0.6667	6
1 (Sleep)	0.6667	0.3333	0.4444	6
Accuracy			0.5833	12
Macro Avg	0.6111	0.5833	0.5556	12
Weighted Avg	0.6111	0.5833	0.5556	12

AUC (macro): 0.5833 APRC (macro): 0.5509

Table 6 GPT-Neo — Zero-Shot Classification Report

	Precision	Recall	F1-Score	Support
0 (Wake)	0.5000	1.0000	0.6667	6
1 (Sleep)	0.0000	0.0000	0.0000	6
Accuracy			0.5000	12
Macro Avg	0.2500	0.5000	0.3333	12
Weighted Avg	0.2500	0.5000	0.3333	12

AUC (macro): 0.5000 APRC (macro): 0.5000

Across all models evaluated under 3-fold cross-validation, GRU achieved the highest accuracy of 75.96% and AUC of 0.7404, while LSTM and RNN performed comparably at 74.60% and 74.62%, respectively. The Transformer achieved the lowest accuracy of 73.90%, suggesting that self attention based mechanisms may require more data or possibly longer sequences to outperform recurrent alternatives. Zero-shot LLMs underperformed specialised models, mirroring recent EEG/ECG findings[5,6,21,22].

Combined ROC Graph for LLMs

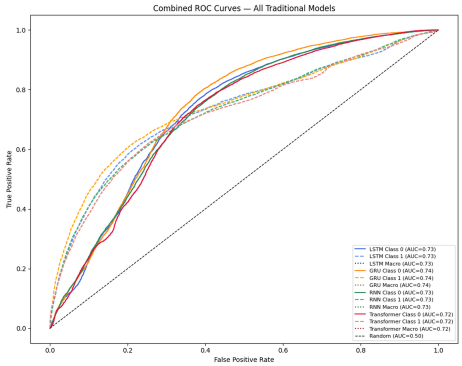

Figure. 6 Combined ROC Curves for Traditional Models

Classification reports were also the main form of tracking data for the LLMs. In their case, tuning was not possible. This also led to the overall ROC curve being a comparison of both models, instead of differing stages within the same training data.

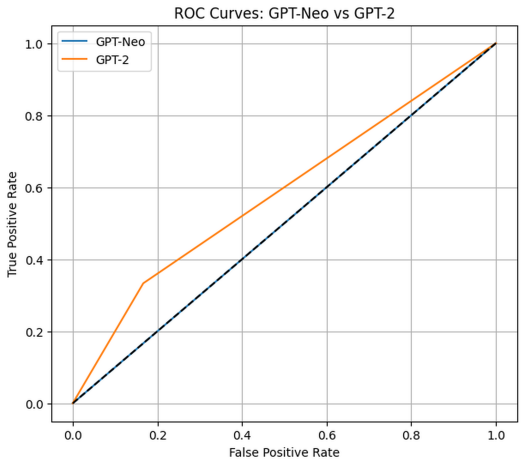

Figure. 7 Combined ROC Curves for Large Language Time Series Models

The following chart shows accuracy per model derived from the average of 3-fold cross-validation results. GRU achieves the highest accuracy among all models at 75.96%, and all of the four deep learning models outperform both zero-shot LLM baselines.

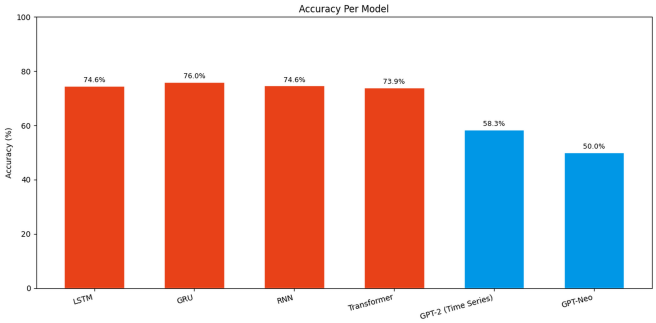

Figure. 8 Combined Accuracies

Discussion

Two trends stand out: (1) all four deep learning models cluster between 73.90% and 75.96% accuracy, substantially above the zero-shot LLM baseline, and (2) GRU achieved the highest accuracy of 75.96% with AUC 0.7404. This outperforms the Transformer, which achieved the lowest accuracy across the four deep learning models at 73.90%. The clustering of recurrent models most likely reflects the suitability of gating mechanisms for capturing short-to-medium range dependencies in EMG signals. The classification visualisations suggest RNNs can overfit local oscillatory patterns without improving the overall metrics.

GRU's strong performance suggests its gating mechanism is well suited for capturing the large amount of short to medium length patterns that are present in EMG sleep signals. On the other hand, the Transformer appears to require longer sequences or larger datasets than those available in this study to demonstrate its advantages, consistent with evidence that attention-based models benefit from richer and more varied input data[9,10].

These accuracies are near but still lower than high-end sleep staging systems which generally achieve 80-85% on multi class classification using EEG[17]. The remaining gap suggests that EMG as a signal by itself might not have enough information for specific staging. Possibly expanding to multimodal inputs such as EEG and EOG could reduce the accuracy gap.

There are 5 major limitations to the study's findings. First, the binary Wake vs Sleep framing is an oversimplification of the clinically relevant 5-class staging problem, and performance on the full multi-class task would likely be different. Secondly, the dataset used in the study suffers from a class imbalance with Wake accounting for 69% of the windows. This imbalance explains the high Wake recall for all models and the low sleep recall as shown in the classification reports. Third, design choices for zero shot LLM evaluation, such as tokenisation, prompt structure, and truncation, have the potential to be detrimental to these models compared to more purpose-built implementations. Next, given the computational cost of rerunning inference on 26,992 windows given context window constraints, LLM baselines were evaluated on the majority aggregate values of the original pipeline's test split. Finally, the restriction of analysis to EMG ignores the more valuable information available from the EEG and EOG channels that are typically used for sleep staging. These limitations preclude strong claims about the absolute performance of the approaches tested, however the relative performance ordering of models is well supported by the cross validated results.

Zero-shot LLMs were not competitive here, which is consistent with reports that unadapted LLMs struggle with raw biosignals and require task-specific encoders, prompts, or multimodal alignment[2,3,5,6]. In ECG, test-time clinical priors can help, but do not eliminate the gap to specialised models[20–22]. Looking forward, multimodal LLM frameworks that fuse EEG/EMG/EOG and behavioural inputs with domain-adapted encoders, or graph-structured time-series models that capture cross-channel dependencies, represent the most promising directions for closing the gap to higher end sleep staging.

When looking towards ROC curves specifically, the curves reinforce these findings. GPT-Neo achieves an AUC of 0.50, sitting directly on the chance diagonal, confirming performance no better than random classification on raw EMG sig-

nals. GPT-2 marginally exceeds chance at AUC 0.58. Among the deep learning models, GRU achieves the highest AUC of 0.7404, followed by LSTM at 0.7302, RNN at 0.7255, and Transformer at 0.7195. The consistent separation between the deep learning curves and the LLM baselines across the full ROC space provides strong visual confirmation that task-specific trained models substantially outperform zero-shot LLMs on this biosignal classification task.

Data Availability

The Sleep-EDF Expanded Cassette dataset is publicly available through PhysioNet[29,30]. In addition, there are existing public program scripts for fitting data to a model. Overall, public availability has made data and its utilization very possible. However, it is not entirely perfect; sleep data, while its labelling and specifications may be vast, don't necessarily have the breadth of patients to reach all possible conditions. This has the potential to force a balance between generalized models and unaccounted sleep circumstances.

Conclusion

This study benchmarked four deep learning architectures and two zero-shot LLM baselines for EMG-based sleep stage classification using the Sleep-EDF Expanded Cassette dataset, evaluated across 3-fold cross-validation. Computational constraints limited the pool of testable LLM architectures and precluded LLM evaluation on the full epoch-level dataset.

The results confirmed the central hypothesis that task-specific deep learning models outperform zero-shot LLMs on biosignal classification. All four deep learning models achieved accuracies between 73.90% and 75.96% across 26,992 windowed samples, well above GPT-2 at 58.33% and GPT-Neo at 50.00%. GRU emerged as the strongest model at 75.96% accuracy and AUC 0.7404, with its compact gating mechanism proving well-suited to capturing temporal EMG patterns. GPT-Neo's chance-level AUC of 0.50 reinforces the broader finding that unadapted language models require domain-specific adaptation before clinical deployment. Future work should explore multimodal inputs combining EEG, EOG, and EMG, full five-class staging, and fine-tuned LLM approaches to bridge the gap between general-purpose language models and task-specific biosignal classification.

Acknowledgments

I would like to thank Munib Mesinovic for his guidance and mentorship throughout this research project.

References

1 A. Sano, J. Amores, M. Czerwinski. Exploration of LLMs, EEG, and behavioral data to measure and support attention and sleep. ArXiv (Cornell University). 2024, https://doi.org/10.48550/arxiv.2408.07822.

2 N. Babu, J. Mathew, V. A. P. Large language models for EEG: a comprehensive survey and taxonomy. ArXiv.org. 2025, https://arxiv.org/abs/2506.06353.

3 N. Chan, F. Parker, W. Bennett, T. Wu, M. Y. Jia, J. Fackler, K. Ghobadi. MedTsLLM: leveraging LLMs for multimodal medical time series analysis. ArXiv (Cornell University). 2024, https://doi.org/10.48550/arxiv.2408.07773.

4 H. Yuan. Letter to the editor. electroencephalography for monitoring brain function and predicting prognosis in patients with TBI. Journal of Neurosurgery. Vol. 141, pg. 876–878, 2024, https://doi.org/10.3171/2024.2.jns24261.

5 J. Kim, A. Alaa, D. Bernardo. EEG-GPT: exploring capabilities of large language models for EEG classification and interpretation a preprint. 2024.

6 D.-H. Lim, M.-J. Cho, H.-H. Kim. Classification of non-invasive EEG signals during motor imagery tasks using a large language model. 2024.

7 H. Liu, X. Liu, D. Yang, Z. Liang, H. Wang, Y. Cui, J. Gu. TodyNet: temporal dynamic graph neural network for multivariate time series classification. ArXiv (Cornell University). 2023, https://doi.org/10.48550/arxiv.2304.05078.

8 S. Hossein Sadat Hosseini, N. N. Joojili, M. Ahmadi. LLMT: a transformer-based multi-modal lower limb human motion prediction model for assistive robotics applications. IEEE Access. Vol. 12, pg. 82730–82741, 2024, https://doi.org/10.1109/access.2024.3413576.

9 W. Fan, J. Fei, D. Guo, K. Yi, X. Song, H. Xiang, H. Ye, M. Li. Towards multi-resolution spatiotemporal graph learning for medical time series classification. pg. 5054–5064, 2025, https://doi.org/10.1145/3696410.3714514.

10 Q. Huang, L. Shen, R. Zhang, S. Ding, B. Wang, Z. Zhou, Y. Wang. CrossGNN: confronting noisy multivariate time series via cross interaction refinement. NeurIPS 2023. 2023, https://openreview.net/forum?id=xOzlW2vUYc.

11 R. N. Sekkal, F. Bereksi-Reguig, D. Ruiz-Fernandez, N. Dib, S. Sekkal. Automatic sleep stage classification: from classical machine learning methods to deep learning. Biomedical Signal Processing and Control. Vol. 77, pg. 103751, 2022, https://www.researchgate.net/publication/360539752_Automatic_sleep_stage_classification_From_classical_machine_learning_methods_to_deep_learning.

12 I. S. Masad, A. Alqudah, S. Qazan. Automatic classification of sleep stages using EEG signals and convolutional neural networks. PLOS ONE. Vol. 19, pg. e0297582, 2024, https://doi.org/10.1371/journal.pone.0297582.

13 T. I. Toma, S. Choi. An end-to-end convolutional recurrent neural network with multi-source data fusion for sleep stage classification. Proceedings of ICAIIC 2023. pg. 564–569, 2023, https://www.researchgate.net/publication/369487737_An_End-to-End_Convolutional_Recurrent_Neural_Network_with_Multi-Source_Data_Fusion_for_Sleep_Stage_Classification.

14 A. Rehman, M. Moussa, H. Saleh, A. Khraibi, A. Khandoker, M. Al-Qutayri. Exploring the role of chin electromyography in automatic sleep stage scoring. Heliyon. Vol. 11, pg. e42122, 2025, https://doi.org/10.1016/j.heliyon.2025.e42122.

15 S. K. Satapathy, B. Brahma, B. P. Biswal, A. K. Bhoi. Machine learning-empowered sleep staging classification using multi-modality signals. BMC Medical Informatics and Decision Making. Vol. 24, 2024, https://doi.org/10.1186/s12911-024-02522-2.

16 L. A. Moctezuma, Y. Suzuki, J. Furuki, M. Molinas, T. Abe. GRU-powered sleep stage classification with permutation-based EEG channel selection. Scientific Reports. Vol. 14, pg. 17952, 2024, https://doi.org/10.1038/s41598-024-68978-4.

17 A. Supratak, H. Dong, C. Wu, Y. Guo. DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Transactions on Neural Systems and Rehabilitation Engineering. Vol. 25, pg. 1998–2008, 2017, https://doi.org/10.1109/tnsre.2017.2721116.

18 M. Perslev, M. Jensen, S. Darkner, P. Jørgen Jennum, C. Igel. U-Time: a fully convolutional network for time series segmentation applied to sleep staging. Neural Information Processing Systems. 2019, https://proceedings.neurips.cc/paper/2019/hash/57bafb2c2dfeefba931bb03a835b1fa9-Abstract.html.

19 M. Yazdi, M. Samaee, D. Massicotte. A review on automated sleep study. Annals of Biomedical Engineering. Vol. 52, pg. 1463–1491, 2024, https://doi.org/10.1007/s10439-024-03486-0.

20 K. Yang, M. Hong, J. Zhang, Y. Luo, S. Zhao, O. Zhang, X. Yu, J. Zhou, L. Yang, P. Zhang, M. Qiao, Z. Nie. ECG-LM: understanding electrocardiogram with a large language model. Health Data Science. Vol. 5, 2025, https://doi.org/10.34133/hds.0221.

21 C. Liu, Z. Wan, C. Ouyang, A. Shah, W. Bai, R. Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. ArXiv (Cornell University). 2024, https://doi.org/10.48550/arxiv.2403.06659.

22 Y. Han, X. Liu, X. Zhang, C. Ding. Foundation models in electrocardiogram: a review. 2024.

23 F. Rivas, J. E. Sierra-Garcia, J. M. Camara. Comparison of LSTM- and GRU-type RNN networks for attention and meditation prediction on raw EEG data from low-cost headsets. Electronics. Vol. 14, pg. 707, 2025, https://doi.org/10.3390/electronics14040707.

24 S. Zhang, Y. Hu, X. Yi, S. Nanayakkara, X. Chen. IntervEEG-LLM: exploring EEG-based multimodal data for customized mental health interventions. pg. 2320–2326, 2025, https://doi.org/10.1145/3701716.3717550.

25 A. Mishra, S. Shukla, J. Torres, J. Gwizdka, S. Roychowdhury. Thought2Text: text generation from EEG signal using large language models (LLMs). 2025.

26 N. Gruver, M. Finzi, S. Qiu, A. G. Wilson. Large language models are zero-shot time series forecasters. Advances in Neural Information Processing Systems. Vol. 36, 2023, https://doi.org/10.48550/arXiv.2310.07820.

27 X. Zhang, R. R. Chowdhury, R. K. Gupta, J. Shang. Large language models for time series: a survey. Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24). pg. 8335–8343, 2024, https://doi.org/10.24963/ijcai.2024/921.

28 A. Anwar, Y. Khalifa, J. L. Coyle, E. Sejdic. Transformers in biosignal analysis: a review. Information Fusion. Vol. 114, pg. 102697, 2025, https://doi.org/10.1016/j.inffus.2024.102697.

29 B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, J. J. L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. IEEE Transactions on Biomedical Engineering. Vol. 47, pg. 1185–1194, 2000, https://doi.org/10.1109/10.867928.

30 A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet. Circulation. Vol. 101, 2000, https://doi.org/10.1161/01.cir.101.23.e215.