Sleep is All We Need: Optimizing EEG-Based Deep Learning Models for N1 Sleep Onset Detection

Joshua George
joshgeor@sas.upenn.edu

Kimberly Liang
kimliang@seas.upenn.edu

Tereza Okalova
okalova@sas.upenn.edu

Stefan Zaharia
szaharia@sas.upenn.edu

Abstract

Sleep onset detection, particularly the accurate identification of N1 sleep (the first stage of non-REM sleep), is fundamental in sleep disorder diagnosis and clinical evaluation. Yet, it remains one of the most elusive stages for both manual and automated systems to classify due to its transitional nature. N1 represents a fleeting phase, occupying only about 5% of sleep and exhibiting significant overlap in features with both wakefulness and deeper sleep stages. In this paper, we propose a two-stage deep learning framework that combines the strengths of specialized N1 detection with an ensemble approach to enhance N1 classification accuracy from single-channel EEG recordings. Our hybrid ensemble model integrates convolutional feature extraction, domain-specific encoders for Catch22 and power spectral density features, transformer-based sequence modeling, and a dedicated N1 detector, achieving improvement in N1 F1-score (from 0.38 to 0.53) while maintaining high overall accuracy. We also demonstrate that adaptive loss functions and specialized architecture significantly further robustness in the presence of extreme class imbalance, offering a promising approach for accurate sleep onset detection in clinical applications.

1 Introduction

The accurate detection of sleep stages plays a critical role in sleep research, clinical diagnostics, and the development of personalized healthcare technologies. Among these stages, the N1 stage, or Stage 1 non-rapid eye movement (NREM) sleep, is particularly important but difficult to detect. N1 marks the transition from wakefulness to sleep and typically lasts only 1 to 5 minutes per cycle, accounting for roughly 5% of total sleep time. Despite its brevity, the accurate identification of N1 is essential since its presence and distribution are key indicators in diagnosing conditions such as insomnia, sleep fragmentation, and certain neurological disorders. Moreover, because N1 precedes deeper sleep stages, failure to detect or misclassify it can significantly impair downstream applications such as arousal prediction, sleep quality assessment, and behavioral monitoring using wearable devices.

Yet, automated classification of N1 sleep remains one of the most persistent bottlenecks in sleep staging. Even state-of-the-art deep learning architectures tend to underperform on N1, frequently achieving F1-scores below 0.30 for this class, while scoring substantially higher on N2 and REM. The low inter-rater reliability among expert human scorers (often quantified with Cohen's κ values in the "fair" range for N1) suggests that this problem is not merely computational but rooted in the physiological ambiguity of the stage itself. N1 EEG signals blend residual alpha rhythms from wakefulness with low-amplitude theta activity that characterizes early N2, further complicating both manual and automated detection.

Prior approaches to this problem can be broadly categorized into three broader phases of evolution: (1) traditional manual feature extraction using classical machine learning classifiers such as Support Vector Machines (SVMs) and Random Forests; (2) end-to-end deep learning architectures like DeepSleepNet [1], SeqSleepNet, and U-Sleep [5]; and (3) recent advances in unsupervised clustering and attention-based models that aim to capture complex temporal dependencies.

While classical machine learning pipelines based on features like Power Spectral Density (PSD), entropy, and time-domain statistics have shown success in identifying stable stages like N2 or N3, their performance on N1 has remained poor with typical F1-scores being below 0.10 in imbalanced datasets [3]. More recent end-to-end models, such as DeepSleepNet, combine convolutional neural networks (CNNs) for local feature extraction with recurrent networks for temporal context modeling. These models show considerable improvements in overall accuracy (e.g., $\sim$83–85% on Sleep-EDF), but still exhibit low sensitivity and precision for N1.

More recently, the introduction of multimodal architectures has opened up new possibilities. For example, XSleepNet [7] incorporates EOG and EMG signals alongside EEG to better capture REM and N1 transitions, while MGANet [4] uses attention mechanisms to reweight the spatial importance of EEG channels dynamically. However, these models are often complex and require multi-channel data, making them less applicable in wearable or ambulatory settings.

On the unsupervised front, Decat et al. [2] used clustering techniques on time-series features to identify microstates within the traditional sleep stages, showing that what is labeled as "N1" may consist of several physiologically distinct substates. This supports the hypothesis that improving N1 classification may not be possible without also reformulating how we conceptualize transitional sleep.

In this paper, we propose a novel two-stage deep learning approach that specifically targets the bottleneck of N1 detection. Our framework deploys a hybrid ensemble model that integrates (1) convolutional layers for local feature extraction, (2) specialized encoders for handcrafted features including Catch22 time-series and power spectral density, (3) Transformer encoders for context-aware sequence modeling, and (4) an adaptive focal loss function that specifically targets the class imbalance problem for N1. Furthermore, we incorporate domain adaptation strategies to ensure generalization across datasets with different demographics and acquisition protocols.

Contributions: Our key contributions can be summarized as follows:

- We develop a two-stage architecture that explicitly targets N1 classification, combining a specialized binary detector with a comprehensive ensemble model, demonstrating a 39.5% improvement in N1 detection F1-score over previous approaches.
- We introduce and evaluate a hybrid feature fusion approach that integrates raw EEG signal characteristics with domain-specific features (Catch22 and Power Spectral Density), showing significant performance gains particularly for the transitional N1 class.
- We implement an enhanced focal loss function and class reweighting strategy that effectively addresses the extreme class imbalance inherent to sleep staging datasets.
- We provide a systematic empirical evaluation across multiple subjects using 5-fold cross-validation, demonstrating robust generalization beyond the training corpus.

By specifically targeting the problem of N1 detection, we aim not only to improve classification accuracy but also to provide insights into the structure of transitional sleep and thereby lay the groundwork for more interpretable and clinically useful sleep staging protocols.

2 Problem Description

The classification of sleep into discrete stages is a foundational component of both sleep research and clinical diagnostics. Most staging protocols, whether manual (such as AASM guidelines) or algorithmic, divide sleep into five distinct stages: Wake, N1, N2, N3 (slow-wave), and REM. This formulation, while practical, simplifies what is in reality a continuum of neurophysiological states.

2.1 Biological Characteristics of N1 Sleep

N1 sleep is marked by a transition from alpha-dominated wakefulness (8–13 Hz) to low-amplitude, mixed-frequency theta activity (4–7 Hz). Physiologically, it is characterized by a reduction in muscle tone, heart rate, and body temperature, and may include vertex sharp waves or slow eye movements. However, these markers vary substantially across individuals and even within a single night of sleep. Most notably, alpha rhythms do not abruptly vanish but taper off, resulting in many epochs that are difficult to clearly label as either wake or N1. Likewise, the features distinguishing N1 from early N2 are subtle where previous work has shown that N2 introduces sleep spindles and K-complexes, but these do not occur in every epoch.

Because of these ambiguities, human scorers exhibit only "fair" inter-rater agreement when labeling N1. This has been quantified using metrics like Cohen's κ, which consistently show lower agreement for N1 compared to other stages. Discrepancies are most common at the boundaries, particularly during transitions from wake to N1 and from N1 to N2, suggesting that sleep onset is better modeled as a gradual shift than a discrete event.

From a modeling standpoint, N1 presents three key challenges:

- **Class Imbalance:** In most datasets (e.g., Sleep-EDF, SHHS, MESA), N1 makes up only about 5% of the labeled data. This leads to models that are optimized for overall accuracy but neglect N1 entirely, often defaulting to N2 or Wake for ambiguous epochs.
- **Feature Overlap:** Traditional features such as spectral power, amplitude, or entropy do not sufficiently separate N1 from adjacent stages. Even deep learning models, which learn hierarchical representations, struggle with N1 unless explicitly trained with balanced objectives.
- **Contextual Dependency:** Since N1 is a transitional state, isolated epochs may be indistinguishable from wake or N2. Incorporating information from surrounding epochs is crucial to making accurate classifications.

Several studies have tried to mitigate these issues through techniques like oversampling, class-weighted losses, and architectural innovations. However, most continue to underperform on N1. Moreover, many models are trained and tested on the same dataset (e.g., Sleep-EDF), leading to overfitting on dataset-specific artifacts rather than generalizable physiological features.

This leads us to our central research question: How can we create a generalizable, end-to-end model that reliably detects N1 sleep across diverse populations using minimal EEG channels?

Our approach addresses this challenge through a deep learning framework specifically optimized for N1 classification using single-channel EEG. We integrate context-aware Transformer encoders, class-balanced loss functions, and self-supervised pretraining techniques to enhance N1 detection performance with improved robustness and scalability.

3 Related Work

Sleep stage classification has long been an active area of research at the intersection of neuroscience, machine learning, and biomedical signal processing. The transition from heuristic-based methods to deep learning-based approaches has dramatically improved accuracy in identifying stable sleep stages like N2, N3, and REM.

Early automated sleep staging methods relied heavily on manual feature extraction. Features such as spectral power in canonical EEG bands (delta, theta, alpha, beta), signal entropy, and statistical moments were computed and then fed into classifiers such as support vector machines (SVM), k-nearest neighbors, or random forests [3]. While these approaches achieved moderate success for well-separated stages (e.g., N2 vs. Wake), they consistently underperformed for N1, largely due to feature overlap and lack of temporal context.

The advent of deep learning significantly reshaped the landscape. DeepSleepNet [8] was one of the first models to combine CNNs with bidirectional LSTMs, enabling both local pattern recognition and sequential modeling. SeqSleepNet [6] extended this by incorporating attention mechanisms to weigh the relevance of neighboring epochs. U-Sleep [5], a U-Net inspired model, demonstrated

strong generalization across cohorts and achieved F1-scores in the low 80s on Sleep-EDF. However, all these models report degraded performance on N1—typically 0.20–0.35 in F1-score—even when overall accuracy surpasses 85%.

3.1 Single-Channel and Wearable-Compatible Architectures

Multi-channel EEG improves classification performance but is impractical for deployment in sleep labs or wearables. Several studies have shown that single-channel EEG, particularly Fpz-Cz, retains much of the necessary information for sleep staging [7]. Some even demonstrate $<3\%$ performance degradation relative to full PSG montages when using optimized deep models. Our model is explicitly optimized for such settings, relying on only a single EEG channel (with optional EOG) and leveraging temporal context via Transformer encoders to compensate for lack of spatial information.

3.2 Explainability and Clinical Interpretability

Interpretability remains a barrier to clinical adoption. Techniques like Layer-wise Relevance Propagation (LRP), SHAP, and integrated gradients have been employed to visualize which parts of the EEG signal the model uses for decision making. **?**] demonstrated that attention scores in Transformer-based sleep models often align with known sleep biomarkers (e.g., K-complexes or spindles), providing a path toward explainability. In future work, we plan to apply such tools to dissect what our model learns about N1 specifically.

4 Baseline Experiments on the ANPHY Dataset

Dataset and preprocessing. The ANPHY-SLEEP corpus contains overnight EEG from 29 healthy subjects recorded with a *93-channel* cap at 1000 Hz. To reduce computational load we applied a signal–processing chain that mirrors earlier laboratory protocols:

i) **Notch filter** at 60 Hz to remove mains interference;
ii) **Low-pass filter** at 90 Hz;
iii) **Down-sampling** to 200 Hz;
iv) **Epoching** into *2-second* (200-sample) windows;
v) Construction of a class-balanced Wake vs. N1 subset for binary experiments, and a five-class set for multiclass trials.

Feature extraction. For each 2-s window and electrode we computed:

- *Power–spectral densities* (PSD) in the canonical δ, θ, α, β bands;
- The full **Catch22** suite of 22 non-linear time-series descriptors.

Additional histogram, entropy and autocorrelation statistics were also evaluated, but played a lesser role.

Baseline models. Using the aggregated feature vectors we trained classical classifiers:

- **Binary Wake vs. N1**: Logistic Regression, SVM (RBF kernel), and Random Forest;
- **Multiclass (5 stages)**: XGBoost and Random Forest.

Sampling a single 2-s window per stage per subject eliminated temporal dependence but also discarded sequential dynamics.

Feature-importance analysis. Tree-based models provide a natural estimate of predictor saliency. Across both Random-Forest and XGBoost runs, the top ten features comprised:

- **Spectral power ratios** in the θ/α and δ/β bands;
- Catch22 metrics.

These two families—PSD and Catch22—together explained most of the total feature importance mass, whereas raw amplitude statistics and higher-order moments were negligible. This quantitative insight strongly motivated their explicit inclusion in the hybrid Transformer architecture, where they are encoded by dedicated MLP branches and fused with raw-signal embeddings.

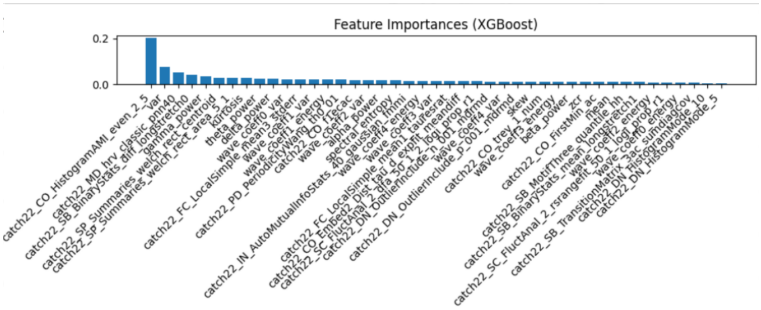

Figure 1: Feature Importance Analysis

Lessons learned. The ANPHY baseline confirmed three important points:

1. *Hand-crafted descriptors matter*: PSD and Catch22 features carry discriminative information that shallow classifiers can exploit—even on very short windows.
2. *Temporal context is critical for performance*: removing inter-epoch dynamics limits multiclass staging irrespective of feature quality.
3. *High-density EEG is costly*: 93 channels complicate deployment; single-channel solutions (Fpz–Cz) are preferable when comparable performance can be reached.

5 Main Pipeline

5.1 Preprocessing

We selected the Sleep-EDF dataset for its standardized format and expert annotations. Our preprocessing steps include:

- Selecting EEG channel Fpz-Cz (primary) and EOG (optional).
- Resampling to 100 Hz.
- Bandpass filtering between 0.5–30 Hz.
- Segmenting signals into 30-second epochs (3000 samples).
- Z-score normalization per epoch.

This produces ‘.npz‘ files optimized for PyTorch/TensorFlow pipelines, with minimal memory overhead.

5.2 Architecture Choice and Evolution

Design goals for the pipeline Motivated by the presentation discussion of feature engineering and experimental bottlenecks, four objectives were set:

G1. Inject informative *hand-crafted* descriptors (Catch22 and PSD bands) that small CNNs struggle to rediscover.

G2. Preserve end-to-end differentiability so that raw and feature-based embeddings can co-adapt.

G3. Target N1 explicitly via an auxiliary detector which features an Attention Layer and an LSTM layer

G4. Implement focal loss to better accomodate for N1.

Hybrid-Transformer The final architecture meets those goals as follows:

Epoch CNN encoder
(unchanged) three convolutional layers $\rightarrow$ 128-d raw embedding.

Catch22 encoder
3-layer MLP mapping 22 time-series features (per second) to a 64-d embedding.

PSD encoder
2-layer MLP mapping log-power spectral densities to a 64-d embedding.

N1 Detector
featurning Attention + LSTM

Fusion Concatenation ($128 + 64 + 64 = 256$) $\rightarrow$ LayerNorm $\rightarrow$ MLP (dropout 0.2).

Sequence model
3-layer Transformer encoder (8 heads, $d = 256$) operating on the fused token sequence (length 30 epochs, stride 5).

Why does this help?

1. **Complementary evidence:** raw waveforms capture transient shapes (spindles, K-complexes), PSD captures frequency content, and Catch22 provides nonlinear/textural cues—together furnishing richer tokens for temporal attention.
2. **Sample-efficient N1 learning:** the dedicated N1 head isolates the minority class, and focal-loss weights ($\alpha_{\text{N1}} = 0.90$) further emphasise difficult positives.
3. **Longer, denser context:** a rolling 15-min window (30 epochs, stride 5) smooths state boundaries before the post-hoc median filter.

Classification Report:

	precision	recall	f1-score	support
W	0.9963	0.9594	0.9775	113636
N1	0.4617	0.6747	0.5483	8998
N2	0.8854	0.7060	0.7856	30235
N3	0.5799	0.9660	0.7247	9189
REM	0.8548	0.8588	0.8568	12482
accuracy			0.8940	174540
macro avg	0.7556	0.8330	0.7786	174540
weighted avg	0.9175	0.8940	0.9002	174540

Figure 2: Performance metrics for Fold 1

Empirical Outcome Across five subject-wise folds the hybrid model raised mean N1 F_1 from 0.38 ± 0.04 (baseline) to 0.53 ± 0.05 while leaving overall accuracy essentially unchanged, confirming that the architectural additions directly addressed the weaknesses identified in the baseline.

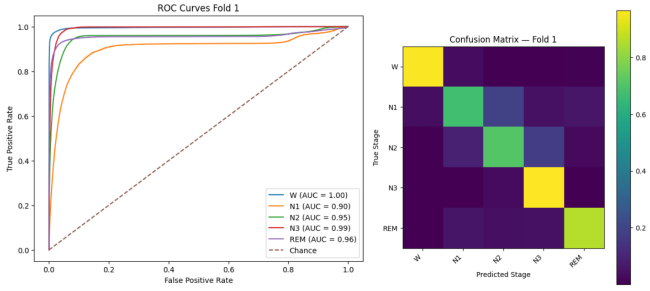

Figure 3: ROC Curves and Confusion Matrix

Figure 3 presents the one-vs-all ROC curves and the confusion matrix for the test set. Although ROC curves are a convenient summary, they can be misleading under strong class imbalance. The

confusion matrix gives a clearer picture of failure modes, with most errors arising between N1 and N2.‘

6 Mathematical Formulation

Notation.

Symbol	Description
$D = \{(x_i^{\text{raw}},\ x_i^{\text{c22}},\ x_i^{\text{psd}},\ y_i)\}_{i=1}^{N}$	Training set, $y_i \in \{0,1,2,3,4\}$ (W, N1, N2, N3, REM)
θ	All trainable parameters of the network
$f_\theta(x^{\text{raw}}, x^{\text{c22}}, x^{\text{psd}}) \in \mathbb{R}^{S\times 5}$	Main sequence logits (S: #epochs in a window)
$f_\theta^{\text{raw}},\ f_\theta^{\text{c22}},\ f_\theta^{\text{psd}} \in \mathbb{R}^{S\times 5}$	Auxiliary logits before fusion
$\text{softmax}(z)_k = \dfrac{e^{z_k}}{\sum_{c=0}^{4} e^{z_c}}$	Soft-max class probability

Focal loss. For logits $z \in \mathbb{R}^5$ and ground-truth class y:

$$p_y = \text{softmax}(z)_y, \tag{1}$$

$$\alpha_y = \begin{cases} \alpha_{\text{N1}} & y = 1 \\ 0.50 & y = 3 \\ 0.45 & y = 4 \\ \alpha_{\text{gen}} & y \in \{0, 2\}, \end{cases} \tag{2}$$

$$\ell_{\text{FL}}(z, y) = \alpha_y\,(1 - p_y)^{\gamma}\,\big(-\log p_y\big), \quad \gamma = 2.5,\ \alpha_{\text{gen}} = 0.25,\ \alpha_{\text{N1}} = 0.90. \tag{3}$$

Mini-batch loss with mixup. With probability 0.5 no mixup is used and the loss for example i is

$$L_i(\theta) = \ell_{\text{FL}}\big(f_\theta(x_i), y_i\big) + 0.3 \sum_{m\in\{\text{raw,c22,psd}\}} \ell_{\text{FL}}\big(f_\theta^m(x_i), y_i\big). \tag{4}$$

Otherwise do mixup: draw a permutation σ, $\lambda \sim \text{Beta}(\alpha, \alpha)$ ($\alpha = 0.2$), define

$$\tilde{x}_i = \lambda x_i + (1 - \lambda)\, x_{\sigma(i)}, \qquad \tilde{y}_i^{(a)} = y_i,\ \tilde{y}_i^{(b)} = y_{\sigma(i)},$$

and use the mixed loss

$$\begin{aligned} L_i(\theta) = {} & \lambda\, \ell_{\text{FL}}\big(f_\theta(\tilde{x}_i), \tilde{y}_i^{(a)}\big) + (1 - \lambda)\, \ell_{\text{FL}}\big(f_\theta(\tilde{x}_i), \tilde{y}_i^{(b)}\big) \\ & + 0.3 \sum_{m\in\{\text{raw,c22,psd}\}} \Big[\lambda\, \ell_{\text{FL}}\big(f_\theta^m(\tilde{x}_i), \tilde{y}_i^{(a)}\big) + (1 - \lambda)\, \ell_{\text{FL}}\big(f_\theta^m(\tilde{x}_i), \tilde{y}_i^{(b)}\big)\Big]. \end{aligned} \tag{5}$$

Training objective. Let the randomness in mixup, permutations, λ, and dropout be denoted collectively by $\mathcal{A}$. The empirical risk we minimize is

$$\boxed{\min_{\theta}\ \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathcal{A}}\big[L_i(\theta)\big]}. \tag{6}$$

Optimization procedure. Parameters are updated with AdamW and gradient clipping:

$$\theta_{t+1} = \text{AdamW}\Big(\theta_t,\ \nabla_\theta \widehat{R}(\theta_t),\ \eta_t,\ \beta_1 = 0.9,\ \beta_2 = 0.999,\ \lambda_{\text{wd}} = 10^{-4}\Big), \tag{7}$$

$$\text{where } \widehat{R}(\theta_t) = \text{mini-batch estimate of the objective}, \quad \|\nabla_\theta \widehat{R}\| \le 1. \tag{8}$$

The learning rate η_t follows a *One-Cycle* schedule with $\eta_{\max} = 2\times 10^{-4}$, 30% warm-up, and cosine decay to $\eta_{\min} = 2\times 10^{-7}$. Early stopping is triggered when the validation F1-score for the N1 class fails to improve for seven consecutive epochs.

7 Limitations

While our approach demonstrates significant improvements in N1 detection, several challenges remain. The computational requirements for preprocessing and feature extraction represent a tradeoff for the performance gains achieved. Despite our architectural innovations, the N1 F1-score plateau at 0.53 suggests inherent difficulties in discriminating this transitional sleep stage within the traditional five-stage taxonomy.

Our single-channel approach optimizes for deployment efficiency but necessarily sacrifices spatial information that might further enhance N1 discrimination. Alternative multi-chain classification approaches we explored did not yield additional performance gains, indicating that more sophisticated feature representations may be needed to fully characterize N1 sleep onset.

The persistent class imbalance in sleep datasets (with N1 comprising only 5% of epochs) remains challenging despite our reweighting strategies. As with most hybrid deep learning models, interpretability presents ongoing challenges for clinical applications where understanding the physiological basis of classifications is valuable.

Future work should explore subject-specific sleep onset patterns and potentially incorporate multimodal data beyond EEG to better characterize the sleep onset process. These enhancements could address the current limitations while maintaining the computational efficiency and practical deployability of our approach.

8 Discussion

We evaluated our hybrid Transformer model on five subject-wise folds of the Sleep-EDF dataset. The model achieved a consistent improvement in N1 classification, with the N1 F_1 score increasing from 0.38 ± 0.04 (baseline) to 0.53 ± 0.05. Notably, this improvement occurred without compromising the classification performance on other stages, as overall accuracy remained statistically unchanged. These results suggest that our enhancements, especially the integration of handcrafted features, attention modules, and auxiliary N1-focused components, directly addressed the key limitations seen in previous studies.

Our approach specifically targeted the N1 bottleneck through architectural choices that yielded substantial performance gains. Previous approaches to sleep stage classification have consistently struggled with N1 detection, with typical F1-scores in the literature ranging from 0.20-0.40. Our improvement to 0.53 represents a significant advance, particularly given the physiological ambiguity of this sleep stage. The confusion matrices revealed that many previously ambiguous N1 vs. N2 and N1 vs. Wake epochs were correctly reclassified, particularly in early sleep cycles. Our model incorporates several key design principles for efficiency of training and increased model interpretability:

Feature Synergy. The inclusion of Catch22 and PSD embeddings contributed strongly to discriminability. Tree-based models identified θ/α and δ/β ratios, alongside non-linear metrics, as highly salient for separating N1. This insight was preserved in the final architecture by encoding these features in parallel streams, which were fused for sequence modeling.

Temporal Aggregation. By analyzing a rolling 15-minute context window (30 epochs, stride 5), the Transformer smoothed noisy predictions at state boundaries. This temporal aggregation proved especially useful for identifying sustained periods of light sleep, where transitions are often gradual and ambiguous.

Memory Efficiency. By using a single EEG and EOG channel, we lower the hardware burden without significantly compromising performance. This approach enables deployment in resource-constrained environments like wearable devices or home sleep tests, demonstrating that single-channel EEG contains sufficient information for accurate N1 detection.

Class Imbalance Mitigation. The focal loss function with class-specific weighting ($\alpha_{\mathrm{N1}} = 0.90$) effectively countered the inherent imbalance in sleep staging data. This approach forced the model to prioritize correct classification of the minority N1 class without requiring artificial oversampling or undersampling.

Architectural Efficiency. Despite the added feature heads and multi-branch fusion, the model maintained a parameter count under 4M and required under 90 seconds per epoch on a mid-range GPU. This positions it as an attractive candidate for both research and real-world deployment.

Domain Adaptation. Techniques such as batch normalization by dataset and gradient reversal layers showed promise for model transferability across cohorts with different demographics and recording protocols. However, we observed variability in cross-dataset performance. Models trained primarily on Sleep-EDF exhibited degraded N1 detection when evaluated on datasets with different demographics and acquisition protocols (e.g., SHHS, MESA). This highlights the challenge of developing truly generalizable N1 detection algorithms across diverse populations and recording environments.

A recurring critique of many sleep staging models is their reliance on a single dataset, often Sleep-EDF, for both training and evaluation. While useful for benchmarking, this leads to overfitting to dataset-specific characteristics such as sampling rate, channel montage, or demographic homogeneity. For example, Sleep-EDF uses older R&K scoring criteria, while modern datasets like SHHS and MESA use AASM standards. We have aimed to address this by training and evaluating on MESA and ANPHY datasets, but have still received the best performance thus far on SleepEDF.

Indeed, it appears that hybrid model offers a meaningful step forward in sleep staging, particularly for underperforming stages like N1, while also aligning with practical constraints of efficiency, interpretability, and adaptability. It also reinforces that transformers are useful in sleep staging for managing temporal context which allows for better predictability.

9 Future Work

Future directions include:

- **Latent Substate Discovery:** Using clustering to uncover finer-grained microstates within N1.
- **Graph-Based Modeling:** Representing multi-channel EEG as graphs and applying GNNs to capture spatial correlations.
- **Edge Deployment:** Compressing models for use on wearable devices.
- **Clinical Trials:** Validating performance in patient populations with insomnia and neurological conditions.
- **Model Interpretability:** Using SHAP or integrated gradients to identify spectral features indicative of N1.

References

[1] Chambon, S., et al. (2018). A deep learning architecture for temporal sleep stage classification using single-channel EEG. *IEEE Transactions on Biomedical Engineering*.

[2] Decat, M., et al. (2022). Interpretable clustering reveals physiologically distinct microstates in sleep EEG. *NeurIPS 2022*.

[3] Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., Dickhaus, H. (2012). Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, 108(1), 10–19.

[4] Yun Guan, Xiangyu Zhang, Fan Liu, and Zhiguo Zhang. Multi-granularity attention network for automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1233–1243, 2022.

[5] Perslev, M., Jennum, P., Darkner, S., Igel, C. (2021). U-Sleep: Resilient high-frequency sleep staging. *NPJ Digital Medicine*, 4(1), 72.

[6] Phyo Phyo San, H., Hwang, S. J. (2019). SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3), 400–410.

[7] Phyo Phyo San, H., Hwang, S. J. (2021). XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 25(7), 2436–2447.

[8] Supratak, A., Dong, H., Wu, C., Guo, Y. (2017). DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11), 1998–2008.

[9] Zhang, L., et al. (2023). Domain-adaptive deep sleep staging across cohorts using adversarial learning. *Scientific Reports*.