

The background of the slide features several horizontal, overlapping EEG waveforms in a light gray color. These waveforms represent brain activity and are spread across the entire width and height of the slide, providing a thematic backdrop for the title.

# **Deep learning for identifying sleep onset from scalp EEG data**

Joshua George, Kimberly Liang, Tereza Okalova, Stefan Zaharia

STAT 4830: Numerical Optimization

The background of the slide features a faint, repeating ECG (heart rate) pattern in a light gray color, spanning the entire width and height of the image.

# **I. What's our problem?**

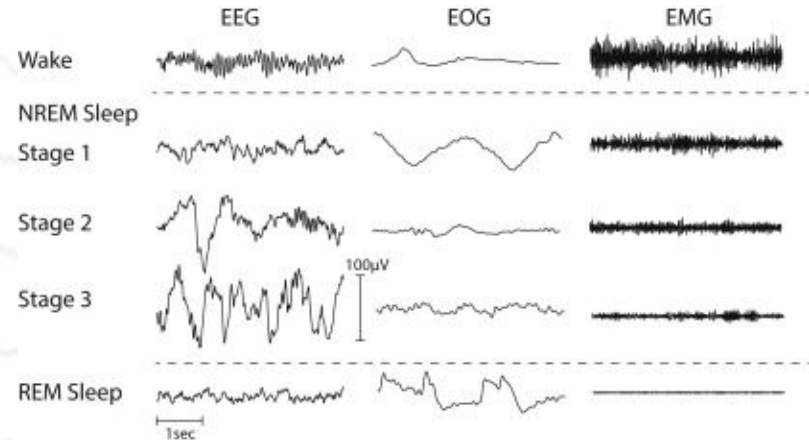
# Unmet need

WHEN WE ALL FALL ASLEEP,  
WHERE DO WE GO?

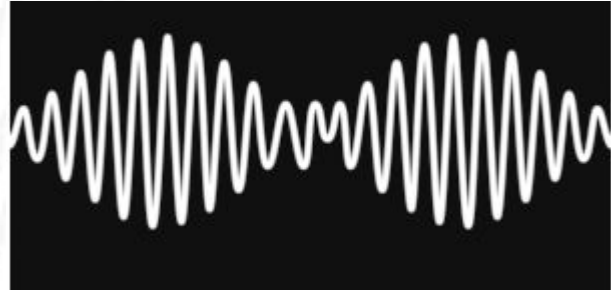
- > 50 million Americans are affected by chronic sleep disorders
- Polysomnography is the gold standard for sleep studies (multimodal)
- Sleep stage scoring labels 30-s PSG epochs
- Manual scoring by experts is **time consuming** and **inconsistent** ( $\kappa$  between human scorers  $\sim .7$ -.8), particularly for identifying sleep onset and stage transitions

# Electroencephalography

- EEG measures electrical activity of the brain using non-invasive scalp electrodes
- Oscillatory brain activity at various frequency bands are collected
- Waveform shape provides insight into different neurophysiological states:
  - Wakefulness
  - Sleep Onset
  - Deep Sleep



# Our favorite bands



$\delta$  (Delta): 0.5-4 Hz — Dominant in N3 (deep sleep)

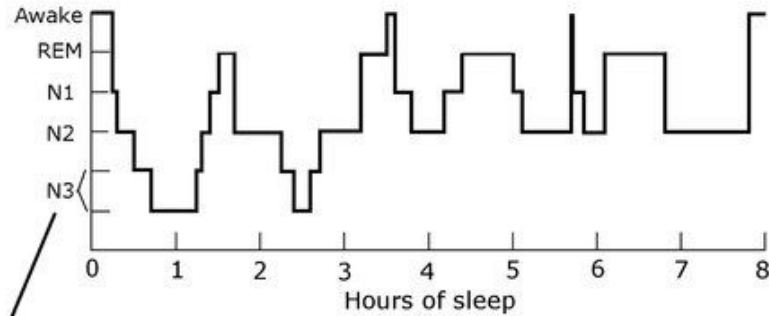
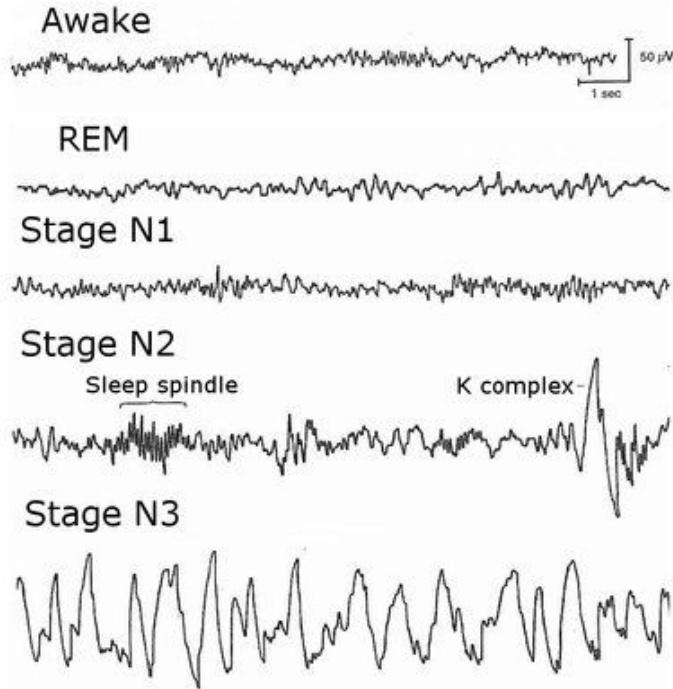
$\theta$  (Theta): 4-8 Hz — Prominent in N1 and REM

$\alpha$  (Alpha): 8-13 Hz — Relaxed wakefulness, attenuates during sleep

$\sigma$  (Sigma): 12-16 Hz — Sleep spindles, characteristic of N2

$\beta$  (Beta): 16-30 Hz — Active wakefulness, cognitive processing

# What makes sleep staging tough



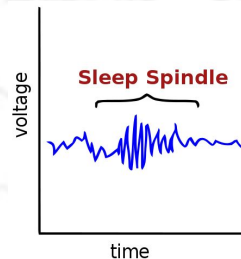
# ...and even tougher

## Why N1 Matters:

- Important for evaluation of sleep efficiency, sleep onset latency, and arousal events.

## Challenges in Clinical Applications:

- Underestimation of wake time can affect patient diagnosis.
- Automated misclassification may lead to inaccurate sleep profiles.

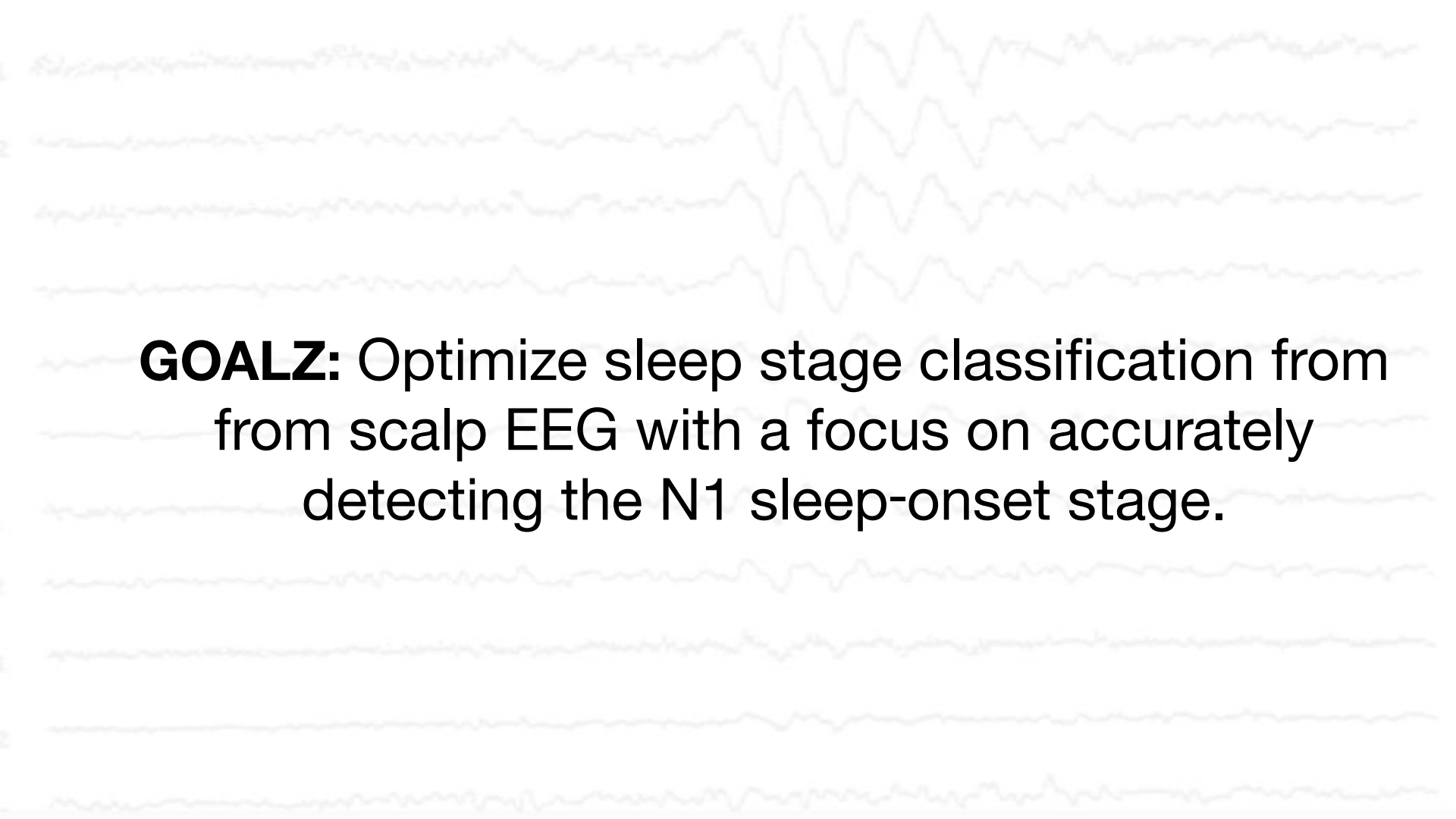


## Transitional Nature:

- N1 is inherently short-lived (typically 1–5 minutes) and transitional.
- Overlaps in physiology with both awake and early N2 states (residual alpha, onset of theta activity).
- absence of spindles makes other validated techniques not work

## EEG Signal Ambiguity:

- residual alpha, onset of theta
- Similarity to relaxed wakefulness and early N2 makes discrimination difficult.
- Add typical N1 duration (1–5 min) and overlap with wake/N2 spectral features.
- Low inter-scorer reliability ( $\kappa < 0.7$ )

The background of the slide features several horizontal EEG waveforms in a light gray color. These waveforms represent brain activity over time, with varying amplitudes and frequencies. The most prominent waveforms are located in the upper half of the slide, while the lower half contains more faint and less distinct traces.

**GOALZ:** Optimize sleep stage classification from scalp EEG with a focus on accurately detecting the N1 sleep-onset stage.



# The (sometimes vicious) cycles

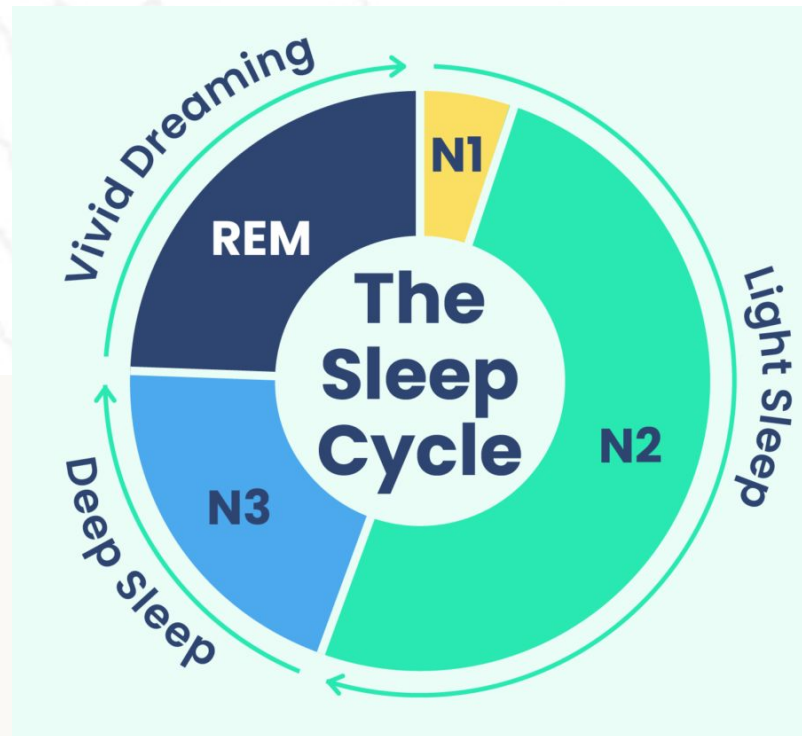
$$f : \mathbf{X} \rightarrow \mathbf{Y}$$

$\mathbf{X} \in \mathbb{R}^{C \times T \times N}$  represents a sequence of  $N$  EEG epochs

- $C$  channels (typically 2: EEG + EOG)
- $T$  time points per epoch (3000 for 30-second epochs at 100 Hz)

•  $\mathbf{Y} \in \{0, 1, 2, 3, 4\}^N$  represents sleep stage labels:

- 0: Wake
- 1: N1 (sleep onset)
- 2: N2
- 3: N3 (deep sleep)
- 4: REM sleep



# Successful Implementation would demonstrate...

- Competitive overall accuracy and F1-scores compared to SOTA classification
- Improved N1 F1-score

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{and} \quad \text{Recall} = \frac{TP}{TP + FN}$$



## **II. What stuff is out there?**

# Traditional Methods & Fully Supervised Settings

- **Time-Series Feature Extraction:**

- Techniques such as spectral power analysis or statistical measures of variance have long been used to extract sleep-relevant features.

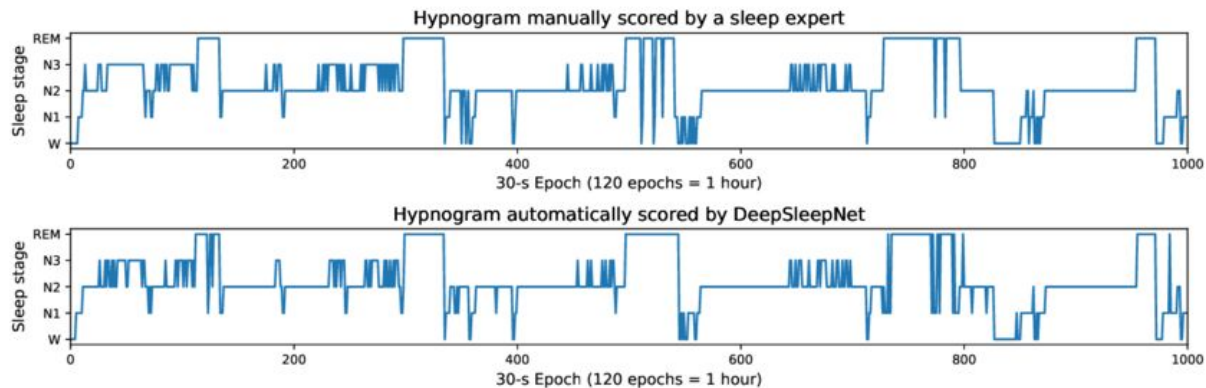
- **YASA (Yet Another (Sleep) Spindle Algorithm):**

- A popular open-source tool that focuses on detecting sleep spindles and other EEG events (e.g., slow waves).
- YASA's reliable detection of spindles and slow-wave patterns can be integrated into feature fusion pipelines, offering context to pure deep learning models.

# Deep Learning Architectures

- **CNN-Based Approaches and Transformers:**

- Early models like DeepSleepNet demonstrated the value of CNNs despite lower N1 performance (F1 around 0.30–0.40)
- Transformer models and related architectures (e.g., SeqSleepNet, U-Sleep) use self-attention mechanisms to model long-range dependencies across epochs.
  - DeepSleepNet (CNN): overall acc = 85 %, N1 F1  $\approx$  0.35



The background of the slide features a series of horizontal, slightly wavy lines in a light gray color. In the center of the slide, there is a more prominent, darker gray waveform that resembles a pulse or a signal, with several sharp peaks and troughs. This waveform is centered vertically and horizontally, creating a focal point for the design.

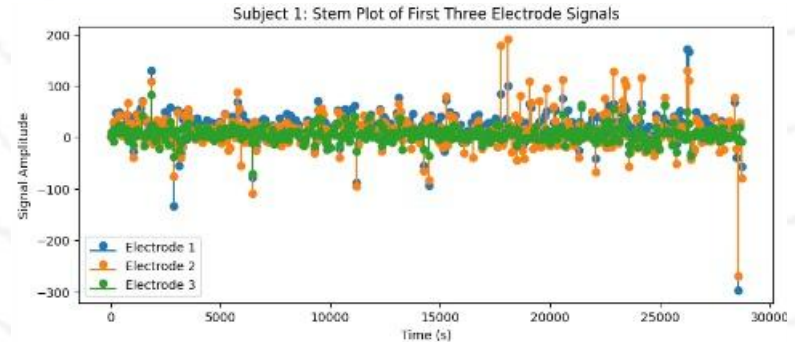
### **III. How supervised methods fared...**

# ANPHY-Sleep dataset without compute

- **Dataset:**
  - ANPHY dataset (29 subjects, 93 EEG electrodes, 1000 Hz sampling rate).
  - Processed into 2-sec EEG windows.
- **Preprocessing:**
  - Notch filtering (60 Hz), low-pass filtering (90 Hz), downsampling (200 Hz).
  - Balanced W vs. N1 dataset.
- **Models:**
  - Binary Classification: Logistic Regression, SVM, Random Forest.
  - Multiclass Classification: XGBoost, Random Forest.
- **Sampling:**
  - Select a fixed time window at random for each stage for each subject
  - Sample a 2-second epoch across all 93 electrodes -> does not preserve dynamical information

# Rationale and pragmatics of using 2-sec windows

- Sleep stage transitions, especially between wake (W) and N1, occur on short timescale
- We hypothesized that 2-second windows would allow us to:
  - Capture rapid neural changes at the W-to-N1 boundary
  - Reduce noise from irrelevant portions of an epoch (since EEG activity fluctuates within longer intervals)
  - Run and iterate quickly





# Spectral features

## Welch's Method for PSD Estimation:

$$P_{xx}(f) = \frac{1}{K} \sum_{i=0}^{K-1} |\mathcal{F}\{x_i(t) \cdot w(t)\}|^2$$

Where:

- $x_i(t)$  is the  $i$ -th segment of the signal
- $w(t)$  is the window function
- $\mathcal{F}$  denotes the Fourier transform
- Parameters: NFFT = 800, window = 400, overlap = 200

## Band Power Calculation:

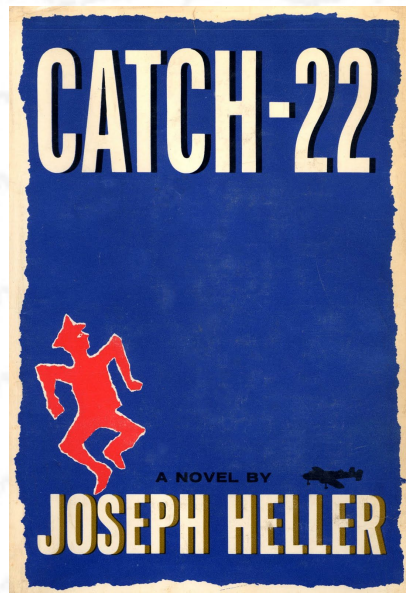
$$P_{band} = \int_{f_{low}}^{f_{high}} P_{xx}(f) df$$

## Spectral Features Computed:

- Band Power Ratios:  $R_{\theta/\alpha} = \frac{P_{\theta}}{P_{\alpha}}$
- Spectral Entropy:  $H_{spect} = -\sum_f \hat{P}_{xx}(f) \log \hat{P}_{xx}(f)$
- Spectral Edge Frequencies: 50%, 90%, 95% of total power

# Some time-domain features of interest

- Modes, outliers, histogram characteristics
- Temporal autocorrelation and stationarity measures
- Information-theoretic measures: entropy and complexity
- Wavelet coefficients and spectral properties
- Transition statistics and motif analysis



**SB\_BinaryStats\_mean\_longstretch1**

**Formula:**

$$b_i = \begin{cases} 1, & x_i > \mu, \\ 0, & \text{otherwise} \end{cases} ; \quad \text{Longest stretch} = \max\{k : b_i, \dots, b_{i+k-1} = 1\}$$

*Intuition:* Measures duration above the mean.

*EEG Relevance:* Detects prolonged high-voltage events (e.g., interictal spikes).

**DN\_OutlierInclude\_p\_001\_mdrmd**

**Formula:**

Identify  $x_i > \mu + 0.001\sigma$ , median inter-event interval.

*Intuition:* Spacing of small upward deviations.

# A couple more...

## CO\_FirstMin\_ac

**Formula:**

$$\tau_{\min} = \min\{\tau > 0 : R(\tau) \text{ local min}\}$$

*Intuition:* Often half the period of dominant oscillation.

*EEG Relevance:* Estimates main rhythm period (e.g., alpha).

## FC\_LocalSimple\_mean3\_stderr

**Formula:**

$$\hat{x}_t = \frac{x_{t-3} + x_{t-2} + x_{t-1}}{3}, \quad \text{RMSE} = \sqrt{\frac{1}{N} \sum_t (x_t - \hat{x}_t)^2}$$

*Intuition:* Local predictability.

*EEG Relevance:* Distinguishes smooth transitions vs. abrupt changes.

## MD\_hrv\_classic\_pnn40

**Formula:**

$$\text{pnn40} = \frac{\#\{|x_{t+1} - x_t| > 0.04\sigma\}}{N - 1}$$

*Intuition:* Proportion of large successive changes.

*EEG Relevance:* Detects abrupt transients or bursts.

## SB\_MotifThree\_quantile\_hh

**Formula:**

$$H = - \sum_{i,j} p_{ij} \ln p_{ij}$$

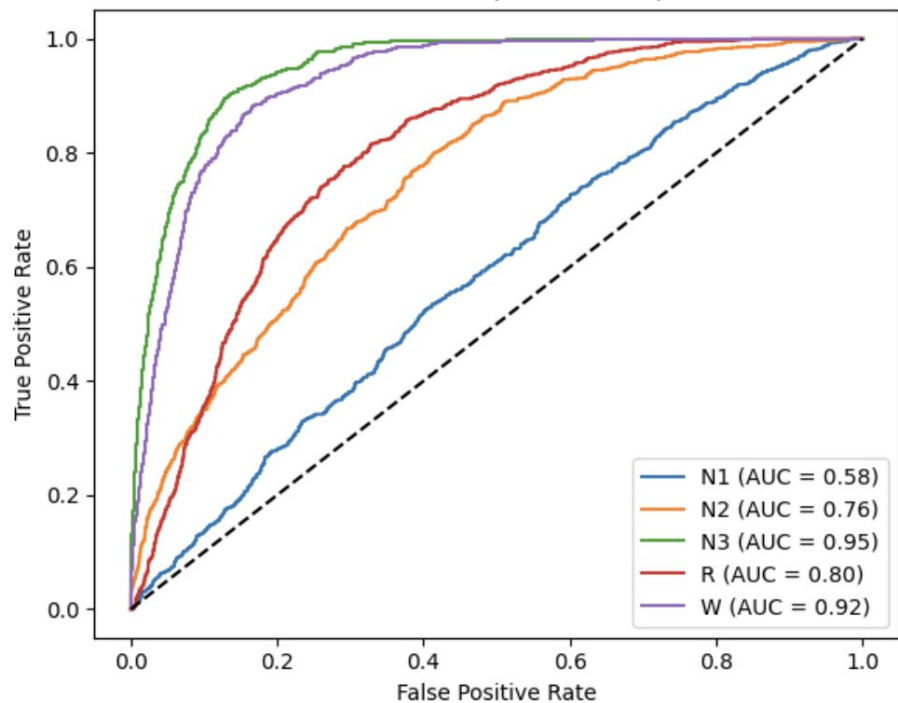
(Symbolize into 3 quantiles; 2-symbol motifs)

*Intuition:* Local pattern complexity.

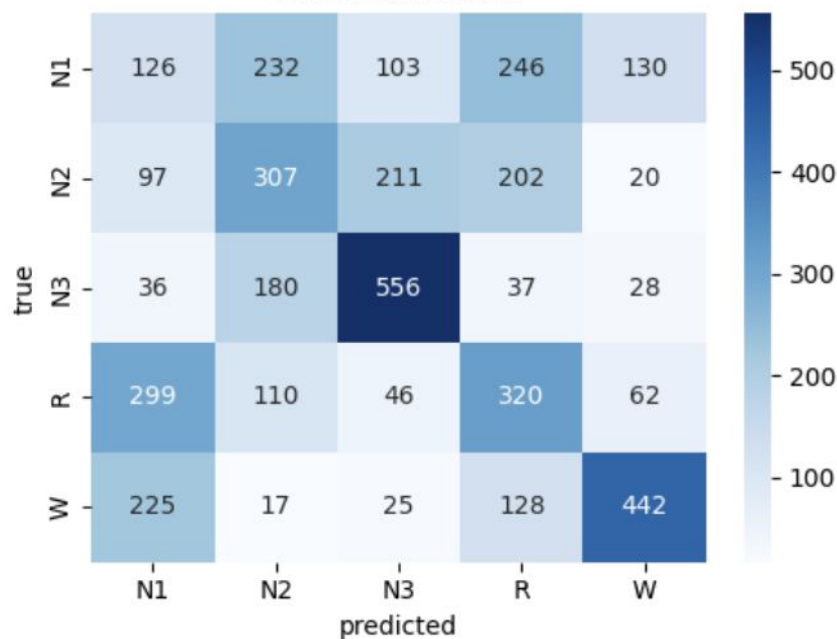
*EEG Relevance:* Higher entropy => richer short-term dynamics.

# Multi-class classification performance

ROC Curves (One-vs-Rest)

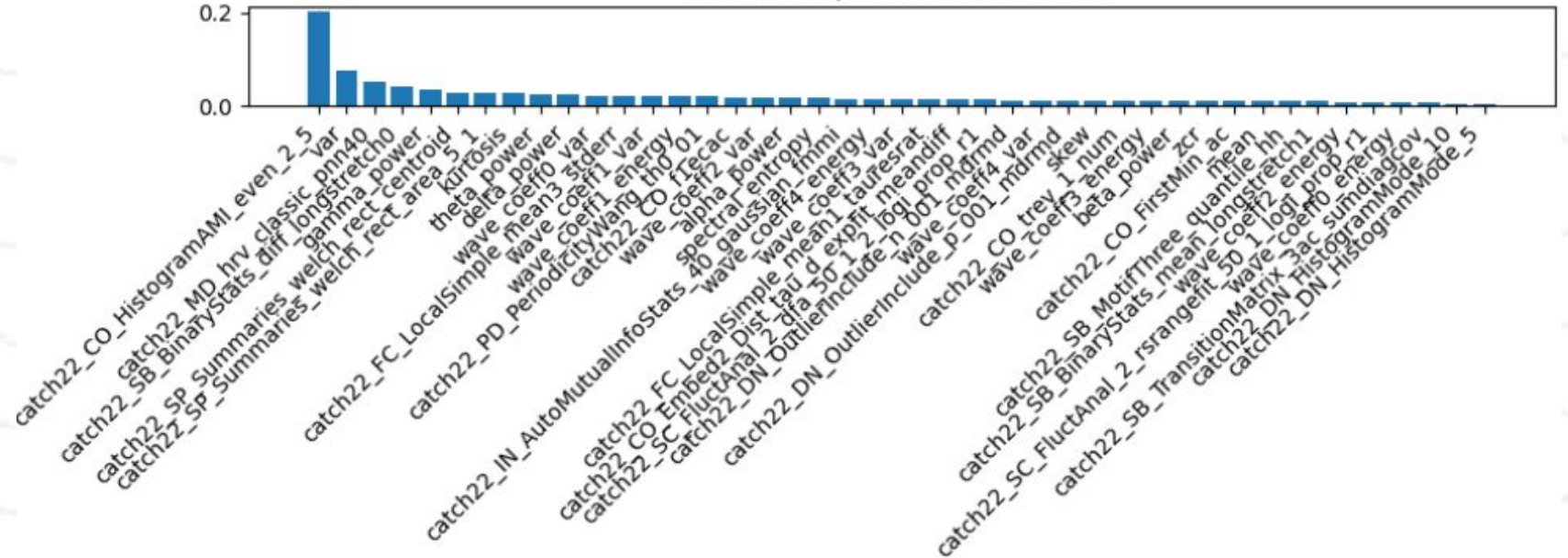


confusion matrix



# Indeed they are useful!

Feature Importances (XGBoost)





## **IV. Hybrid model territory**

# New dataset: Sleep-EDF

PhysioNet repository of polysomnographic sleep recordings (old but ubiquitous)

197 whole-night recordings from 100 subjects

- Sleep Cassette (SC): 153 recordings from 78 subjects (home recordings)
- Sleep Telemetry (ST): 44 recordings from 22 subjects (hospital recordings)

Signals:

- EEG (Fpz-Cz and Pz-Oz channels; currently picking only the former)
- EOG (horizontal eye movements)
- EMG (ignoring for now)

W: 507994 samples (62.61%)

N1: 42712 samples (5.26%)

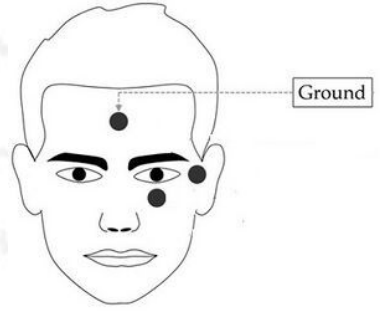
N2: 160147 samples (19.74%)

N3: 38893 samples (4.79%)

REM: 61594 samples (7.59%)



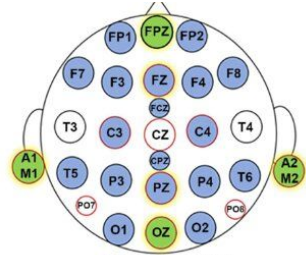
# Preprocessing and raw signal processing



- Temporal filtering:  $x_{filtered}(t) = (h * x)(t)$  where  $h$  is a bandpass filter (0.5-30.0 Hz)
- Normalization:  $\hat{x}(t) = \frac{x(t) - \mu_x}{\sigma_x}$
- Segmentation: 30-second epochs with sequence length 20, stride 10

$$\mathbf{X}_{raw} \in \mathbb{R}^{B \times S \times C \times T}$$

Where  $B$  is batch size,  $S$  is sequence length,  $C = 2$  channels,  $T$  is time points





# Model architecture (surely the neural net doesn't need our help)

## CNN Encoder:

- Input: Sequence of 20 consecutive EEG epochs (2 channels  $\times$  T time points)
- 3 convolutional layers (kernel sizes 5 $\rightarrow$ 3 $\rightarrow$ 3) with ReLU activation and max pooling
- Output: 128-dimensional embedding per 30s epoch

## Transformer:

- Input: Sequence of 20 epoch embeddings
- 2 transformer encoder layers with 4 attention heads (dim = 128)
- Positional encoding (20 $\times$ 128 tensor) for temporal context

## Classification:

- Linear layer maps each transformed embedding to 5 sleep stage probabilities
- Focal loss to address class imbalance in sleep stages
- Median filter (kernel size 5) applied to predictions for temporal consistency
- Special attention to N1 stage detection (typically underrepresented)



# Model architecture (cont'd)

## A. Raw Signal Encoder (CNN):

$$h_i^{(l+1)} = \text{Pool}(\text{ReLU}(\text{BN}(W_i^{(l)} * h_i^{(l)} + b_i^{(l)})))$$

$$\mathbf{h}_{CNN} \in \mathbb{R}^{B \times S \times 128}$$

## \*\*B. Feature Encoders (MLP):\*\*

$$\mathbf{h}_{C22} = \text{MLP}_{C22}(\mathbf{X}_{C22}) \in \mathbb{R}^{B \times S \times 64}$$

$$\mathbf{h}_{PSD} = \text{MLP}_{PSD}(\mathbf{X}_{PSD}) \in \mathbb{R}^{B \times S \times 64}$$

## \*\*C. Feature Fusion:\*\*

$$\mathbf{h}_{fused} = [\mathbf{h}_{CNN}; \mathbf{h}_{C22}; \mathbf{h}_{PSD}] + \mathbf{PE} \in \mathbb{R}^{B \times S \times 256}$$

Where  $\mathbf{PE}$  is a learnable positional encoding

# Model architecture (cont'd)

**\*\*D. Transformer Encoder:\*\***

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O$$

$$\mathbf{z} = \text{TransformerEncoder}(\mathbf{h}_{\text{fused}}) \in \mathbb{R}^{B \times S \times 256}$$

**E. Classification Heads:**

$$\hat{\mathbf{y}} = \text{softmax}(W_c \cdot \mathbf{z} + b_c) \in \mathbb{R}^{B \times S \times 5}$$

# Choices and parameters

- Focal Loss ( $\alpha=0.25$ ,  $\gamma=2$ ) to address class imbalance
  - General classes:  $\alpha = 0.25$
  - N1 (sleep onset):  $\alpha = 0.9$  (much higher)
  - N3:  $\alpha = 0.5$
  - REM:  $\alpha = 0.45$
- gamma set to 2.5 (stronger focus on hard-to-classify examples compared to standard cross-entropy loss)

$$\mathcal{L}_{focal}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$$

- ReduceLROnPlateau (factor 0.5, patience 3)
- WeightedRandomSampler balances classes per batch

$$\text{Class weighting: } w_c = \frac{1}{\sqrt{n_c + \epsilon}}$$

$$w_{N1} = 1.5 \times w_{N1} \text{ (additional N1 boosting)}$$

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j, \lambda \sim \text{Beta}(0.2, 0.2)$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j$$

Mixup for 50% of batches: EEG/EOG signals are combined using a weighted average. Since  $\alpha=0.2$  is small, the Beta distribution creates  $\lambda$  values mostly near 0 or 1, resulting in "gentle blending" where one signal dominates.

# Hyperparameters

- Learning rate:  $2 \times 10^{-4}$
  - Batch size: 32
  - Sequence length: 30
  - Sequence stride: 5
  - Transformer layers: 3
  - Attention heads: 8
  - Dropout: 0.2
- Using a learning rate scheduler helped slightly for N1-focused performance, but a scheduler on cosine loss wasn't as useful
  - Grid search didn't significantly improve class-level performance overall (only minor improvement)
  - Model became unstable after 3 transformer layers — 2–3 layers seemed to work best

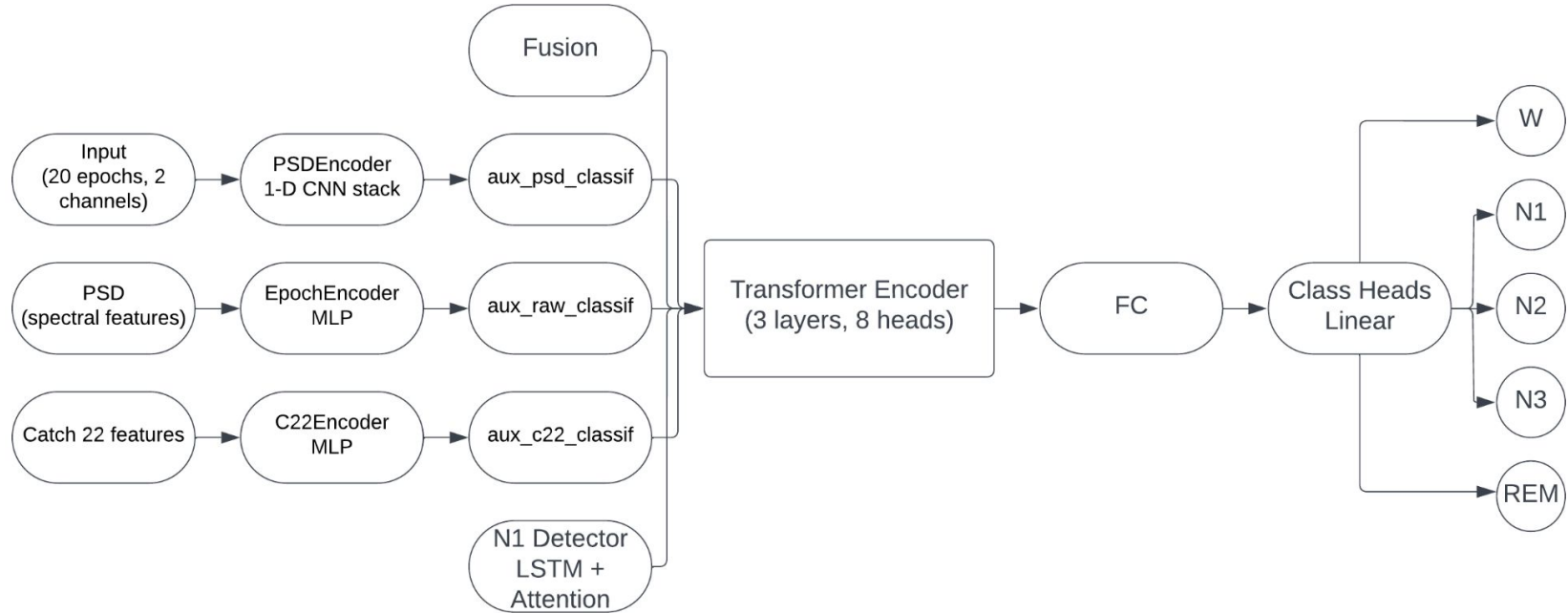
AdamW with weight decay  $\lambda = 1 \times 10^{-4}$

Learning rate  $\eta = 2 \times 10^{-4}$

Gradient clipping:  $\hat{g} = g \cdot \min\left(1, \frac{1.0}{\|g\|_2}\right)$

$$\Theta = \{1e^{-4}, 2e^{-4}, 3e^{-4}\} \times \{32, 64\} \times \{20, 30\} \times \{5, 10\} \times \{2, 3\} \times \{4, 8\} \times \{0.1, 0.2\}$$

# Best of both worlds? (hybrid transformer with pre-computed features)



# Handcrafted features

## PSD:

$$P_{xx}(f) = \frac{1}{K} \sum_{i=0}^{K-1} |\mathcal{F}\{x_i(t) \cdot w(t)\}|^2$$

- Band powers:  $P_\delta, P_\theta, P_\alpha, P_\sigma, P_\beta$
- Spectral ratios:  $R_{\theta/\alpha} = \frac{P_\theta}{P_\alpha}$
- Spectral entropy:  $H_{spect} = -\sum_f \hat{P}_{xx}(f) \log \hat{P}_{xx}(f)$

## Good old Catch22:

$$\phi_{C22} : \mathbb{R}^T \rightarrow \mathbb{R}^{22}$$

$$\mathbf{X}_{C22} \in \mathbb{R}^{B \times S \times 44}$$

$$\mathcal{D} = \{(X_i^{\text{raw}}, X_i^{\text{c22}}, X_i^{\text{psd}}, y_i)\}_{i=1}^N,$$

$$\theta = \{\theta_e, \theta_c, \theta_p, \theta_f, \theta_t, \theta_n, \theta_s, \{\theta_{h_j}\}_{j=0}^4\}.$$

1. EpochEncoder:

$$h_i^r = \text{Enc}^r(X_i^{\text{raw}}; \theta_e) \in \mathbb{R}^{256}$$

2. C22Encoder:

$$h_i^c = \text{Enc}^c(X_i^{\text{c22}}; \theta_c) \in \mathbb{R}^{256}$$

3. PSDEncoder:

$$h_i^p = \text{Enc}^p(X_i^{\text{psd}}; \theta_p) \in \mathbb{R}^{256}$$

4. Fusion + PosEnc:

$$z_i = \text{PosEnc}(\text{Fuse}([h_i^r, h_i^c, h_i^p]; \theta_f)) \in \mathbb{R}^{S \times D},$$

where  $D = 128 + 64 + 64$ ,  $S$  is sequence length.



**5. Transformer:**

$$u_i = \text{Trans}(z_i; \theta_t) \in \mathbb{R}^{(S+5) \times D}.$$

$$\text{Split } u_i = [u_i^{\text{seq}}; u_i^{\text{cls}}].$$

**6. N1 Detector:**

$$d_i = \text{N1Det}(u_i^{\text{seq}}; \theta_n) \in \mathbb{R}^S.$$

**7. Shared FC:**

$$s_i = \text{SharedFC}(u_i^{\text{seq}}; \theta_s) \in \mathbb{R}^{S \times H}, \quad H = 256.$$

**8. Per-class heads (one-vs-all logits):**

$$l_{i,j} = \text{Head}_j(s_i; \theta_{h_j}) \in \mathbb{R}^S, \quad j = 0, \dots, 4.$$

Then **augment** the N1 logits:

$$\ell_{i,1} = l_{i,1} + d_i, \quad \ell_{i,j} = l_{i,j} \quad (j \neq 1).$$

$$\ell_i = [\ell_{i,0}, \dots, \ell_{i,4}] \in \mathbb{R}^{S \times 5}.$$

$$\text{FL}(\ell_i, y_i) = \frac{1}{S} \sum_{t=1}^S \alpha_{y_{i,t}} (1 - p_{i,t,y_{i,t}})^\gamma [-\log p_{i,t,y_{i,t}}],$$

$$\text{FL}_i^r = \text{FL}(\text{Aux}^r(z_i), y_i), \quad \text{FL}_i^c = \text{FL}(\text{Aux}^c(z_i), y_i), \quad \text{FL}_i^p = \text{FL}(\text{Aux}^p(z_i), y_i).$$

The optimization problem:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N \left[ \underbrace{\text{FL}(\ell_i, y_i)}_{\text{main loss}} + \lambda_{\text{aux}} (\text{FL}_i^r + \text{FL}_i^c + \text{FL}_i^p) \right] + \frac{\lambda_2}{2} \|\theta\|_2^2.$$

The background of the slide features a series of horizontal ECG (heart rate) waveforms in a light gray color, spanning the entire width and height of the image. These waveforms are typical of a medical monitor display, showing rhythmic peaks and troughs.

## **V. Results**

# CNN-Transformer (focal loss proportional to samples)

Final Metrics (with smoothing):					
	precision	recall	f1-score	support	
W	0.99	0.95	0.97	110105	
N1	0.39	0.30	0.34	7996	
N2	0.81	0.79	0.80	33877	
N3	0.71	0.79	0.75	8882	
REM	0.63	0.89	0.74	13160	
accuracy			0.88	174020	
macro avg	0.71	0.74	0.72	174020	
weighted avg	0.88	0.88	0.88	174020	

Input  
(20 epochs,  
2 channels)

Epoch Encoder  
(Conv1D×3,  
Pool→FC→128)

Add Positional  
Encoding

Transformer  
Encoder  
(2 layers,  
4 heads)

Output Head  
(Linear→5 classes)

# Same, but with focal loss focused on N1

Epoch 35/35

Train Loss: 0.0769, Acc: 0.7706

Val Loss: 0.0523, Acc: 0.8548

Detailed Metrics:

	precision	recall	f1-score	support
W	0.99	0.92	0.96	110105
N1	0.30	0.57	0.39	7996
N2	0.85	0.70	0.77	33877
N3	0.71	0.82	0.76	8882
REM	0.64	0.88	0.74	13160
accuracy			0.85	174020
macro avg	0.70	0.78	0.72	174020
weighted avg	0.89	0.85	0.87	174020

Input  
(20 epochs,  
2 channels)

Epoch Encoder  
(Conv1D×3,  
Pool→FC→128)

Add Positional  
Encoding

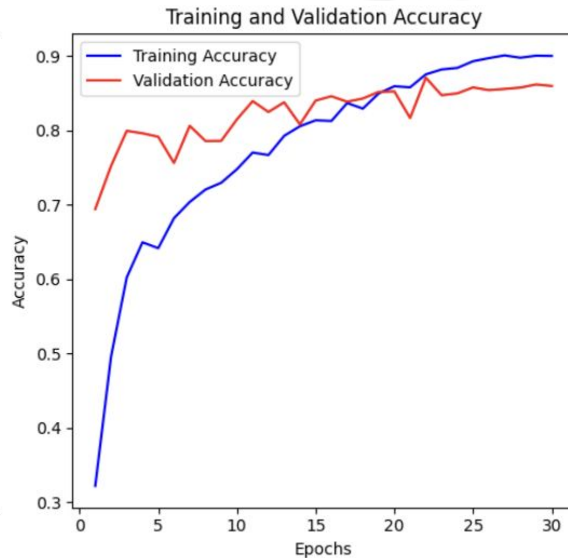
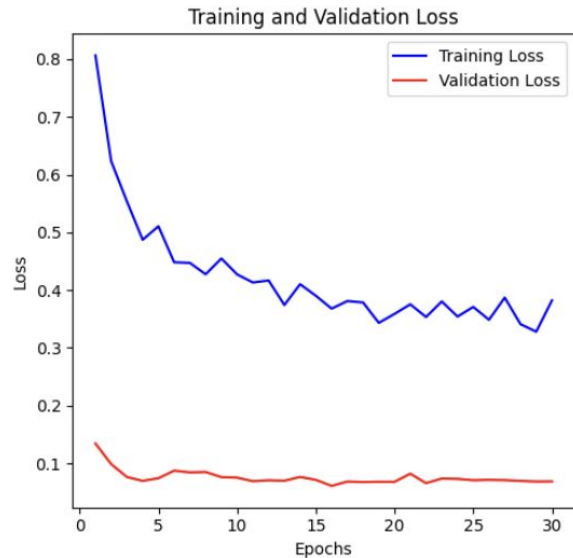
Transformer  
Encoder  
(2 layers,  
4 heads)

Output Head  
(Linear→5 classes)

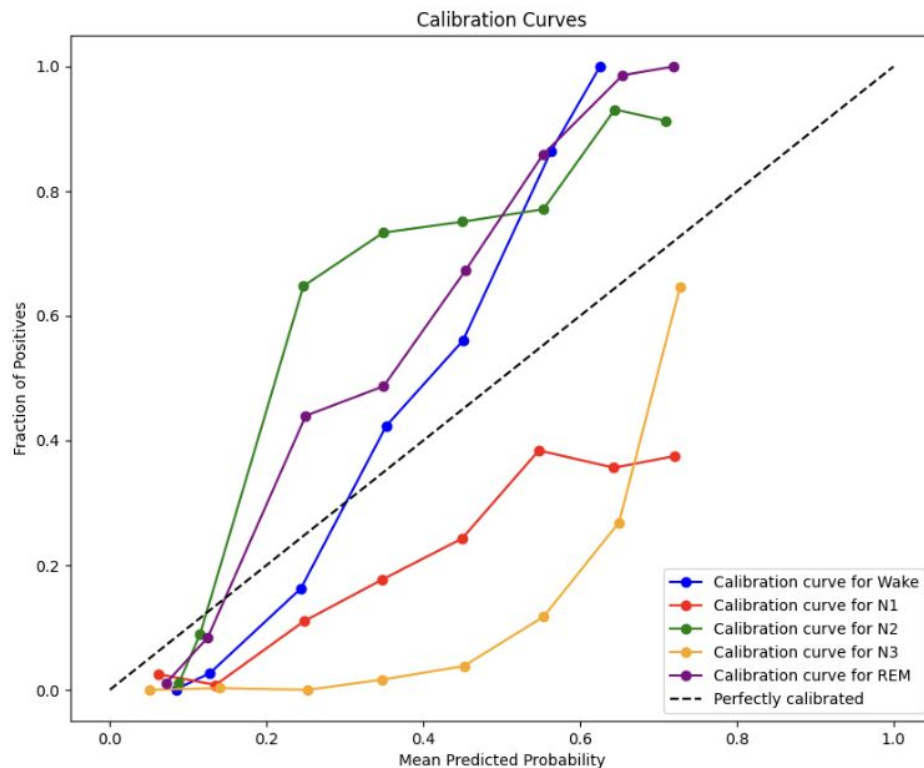
# Hybrid model performance

Metric	Value
Overall Accuracy (Raw)	90.32%
Overall Accuracy (Smoothed)	88.18%
F1-Wake	0.9721
F1-N1 (Sleep Onset)	0.5051
F1-N2	0.7785
F1-N3	0.8040
F1-REM	0.8340
Best Model Epoch	30
Training Stopped At	37 epochs

# Training and validation curves (hybrid model)



# Calibration curves (hybrid model)



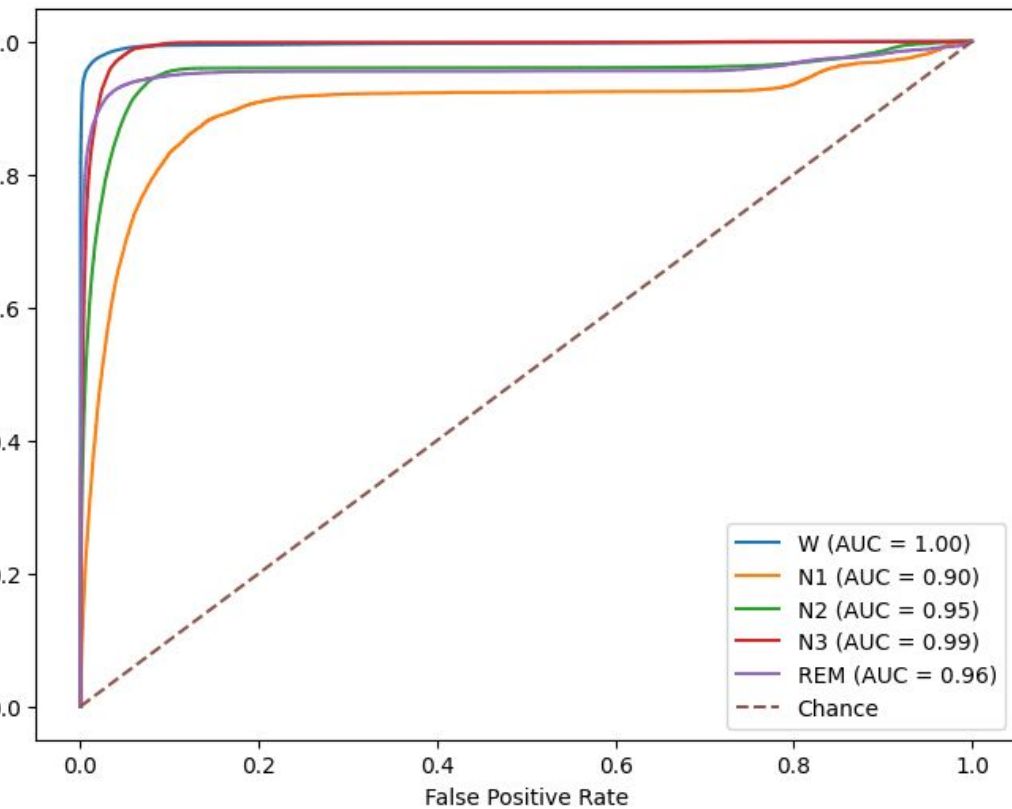


# Lucky fold

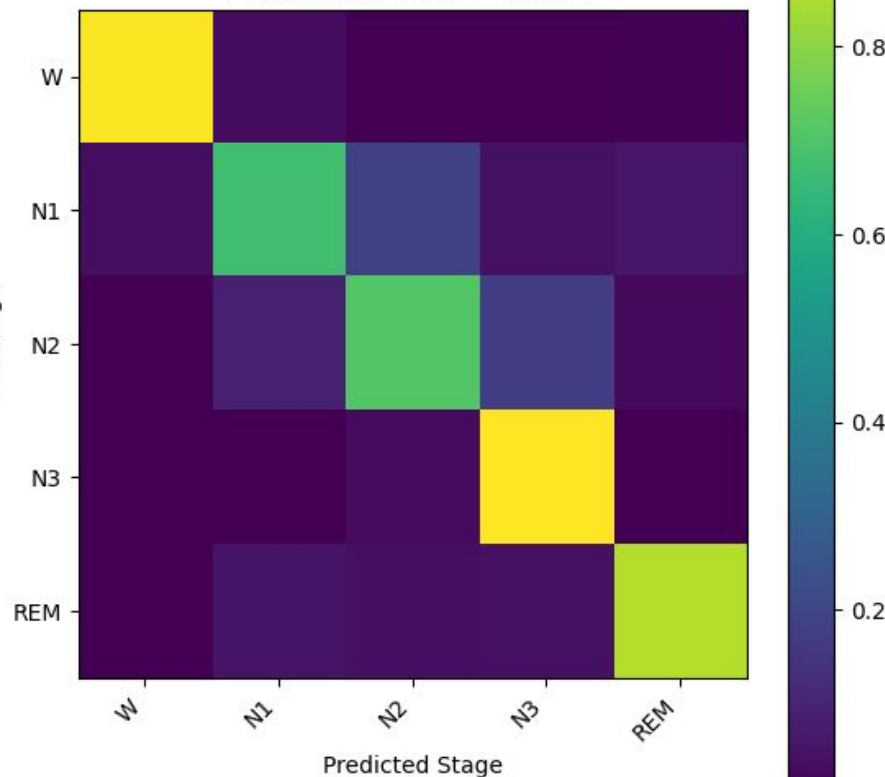
## Classification Report:

	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

ROC Curves Fold 1



Confusion Matrix — Fold 1



The background of the slide features a series of horizontal, slightly wavy lines in a light gray color. In the center of the slide, there is a prominent, darker gray pulse waveform, similar to an ECG or heart rate monitor trace, which spans across the width of the slide and overlaps with the horizontal lines.

## **VI. Biggest Challenges**

- Compute - ran on Colab, local computers, took turns training model
- Storage of datasets
- Data processing speed
- Consistency of procedures across sets
- Integration of additional physiological tools

The background of the slide features a faint, light gray ECG (heart rate) pattern that spans the entire width and height of the image. The pattern consists of multiple horizontal lines with periodic peaks and troughs, resembling a standard medical waveform.

## **VII. Discussion**

## **Key Challenges in Sleep Staging:**

- N1 remains the most challenging label to classify due to its transient and overlapping EEG features with wakefulness and N2.
- Significant class imbalance
- We might be missing out on network-level dynamics by not explicitly considering bivariate/multivariate features

## **Persisting bottlenecks & SOTA:**

- CNN-based architectures (DeepSleepNet, U-Sleep) and Transformer models have improved overall staging but still show suboptimal performance for N1 ( $F1 < 0.5$  in many studies).
- Will ML models remain only as good as the human scorers?

# Looking ahead: utility for intracranial data?

- iEEG offers higher SNR and captures electrical activity directly from cortical and subcortical regions.
- Patients with implanted devices (e.g., for epilepsy or deep brain stimulation) already provide access to high-quality intracranial recordings.
- Using iEEG for sleep staging eliminates the need for extra sensors and enables continuous, real-time monitoring.
- considerably higher sampling rate (~5 kHz), which may improve the delineation of challenging stages (like N1) by resolving ambiguous signals
- treatment management for neurological and psychiatric disorders (e.g. seizures)

# Prediction error? (expected vs. actual progress)

## Expected Progress:

- Train a baseline model in a reasonable amount of time, feel good about life
- See decent initial accuracy and hopefully understand what is happening in our dataset, maybe plot a few nice graphs
- Match and exceed SOTA benchmarks

## Actual progress:

- Spent an alarming amount of time fighting the data (labels refusing to line up, missing values, confusion about patients)
- Ran basic models that were done in literature
- Eventually managed to train a baseline model... only to realize the results were suspiciously good
- Felt paranoia about data leakage.. developed trust issues with validation splits
- Got a good model working eventually once we combined catch22 and power spectral density into the transformer
- Kept iterating until we got  $>0.5$  F1 in N1 sleep staging





**Thank you!!**

# Reflections - Stefan

#1 Technical difficulty: Colab

Easiest part of workflow: coming up with ideas; hardest: implementing + training

Goals: became SURVIVAL

AI tools: writing especially, very helpful with the mathematical formulation

Most surprising result:  $W$  being  $\sim 1$

Most useful lecture topic: Transformer

Optimization in practice: a lot of trial and error, not much math :)

If we had two more weeks: add more data, reduce the size of  $W$  labels

Start again: nothing? I did not consider anything wasteful - good progression. Midway however: should have gotten the patient split right

# Reflections - Kimberly

#1 Technical difficulty: Pioneer/GPU cluster kept disconnecting

1. Realized how hard N1 classification is even reaching 30% accuracy was a major challenge and was surprised that we were able to
2. The Transformers demo really helped make the model architecture feel less confusing and early lectures on computational cost gave a strong foundation for optimizing code later.
3. Learned that optimization is mostly about patience and catching small errors early, not just tuning models.
4. If we had two more weeks: would finish training on other datasets and add spindle detection.
5. If we restarted: would plot outputs earlier and change only one thing at a time instead of mixing experiments.

# Reflection - Tereza

## #1 Technical difficulty: twinning with Kim (Pioneer pain)

1. struggled striking the balance between depth of knowledge of the subject matter versus treating sleep staging as an engineering problem
2. bad intuition about computational load made many of my efforts not realizable
3. Lectures on parallelization filled a huge knowledge gap in my brain and helped me optimize our code
4. Being called out (in a good way) for tweaking too many things at once helped me pivot and make tractable, incremental progress at an ultimately faster rate
5. If we had too more weeks: we would include self-supervised step and more validation data
6. If we could start over: keep a systematic log of our roadmap and intended architecture choices so that it does not feel like a random walk through the parameter space

# Reflections - Joshua

1. Coming in without a good idea about the challenges surrounding N1 classification, it was surprising to see how difficult it was to get a reasonable detection, let alone improved results
2. Getting a formal workflow for tuning DL models was helpful as a way to slow down our process, particularly to combat the desire to keep advancing quickly and obtain more results
3. Certainly I would say having a better plan of attack from the outset, emphasizing organization and consistency in how we address a particular problem. Taking it one problem at a time rather than hoping multiple changes at once will fix our issues. Also, having a better plan for addressing compute
4. Likely comparison with other sets, like SHHS and MESA which we begun, but required more time due to the differing practices across datasets
5. Having better understanding of which datasets to use and how we wanted to process them to obtain consistent outcomes