Maastricht University

Department of Advanced Computing Sciences

Automatic Sleep Disorder Classification

Semester Research Project Final Report

Group 13

Aedem Bangerter
Daniel Skuczi
Eduard Levinschi
Yannick Brackelaire
Jonas Daukša
Stefan Popa

Supervisors:
Dr. Iris Huijben
Dr. Pietro Bonizzi

Coordinator:
Dr. Menica Dibenedetto

January 23, 2026

Abstract

Sleep disorders affect more than one-third of the world's population and contribute to cardiovascular disease and cognitive decline. Traditional diagnosis relies on manual expert scoring of polysomnography recordings, which is time-consuming and varies between scorers. This project tests whether pre-trained foundation models can classify sleep disorders in small, imbalanced clinical datasets.

We evaluated SleepFM, a multimodal foundation model trained on more than 100,000 hours of PSG data, on the CAP Sleep Database that contains 82 full-night recordings in 5 disorder classes. We freeze the pre-trained SleepFM encoders and train simple classifiers on extracted embeddings, comparing against classical machine learning baselines using handcrafted statistical features.

Random Forest trained on SleepFM embeddings achieved the best performance (Macro-F1: 0.41, AUROC: 0.87, AUPRC: 0.71) without special handling of class imbalance. Baseline models needed careful imbalance handling, with adaptive noise oversampling giving the best results (RF: Macro-F1 0.44, MLP: Macro-F1 0.46). SleepFM embeddings appear to have better class separability, reducing the need for heavy rebalancing while maintaining competitive performance.

Contents

1 Introduction

Sleep disorders represent a major health problem that affects more than one-third of the world's population[20]. These conditions are linked to cardiovascular disease, metabolic dysfunction, depression, and cognitive decline. Accurate diagnosis is essential for treatment, yet current practice relies heavily on manual polysomnography (PSG) scoring by trained experts, a labor-intensive process that is time-consuming and inconsistent, with disagreement rates reaching 19.3% across sleep centers[13].

Machine learning in sleep medicine has evolved from statistical features to deep learning architectures for automatic sleep stage classification. Recent models like U-Sleep[22] and SleepTransformer[23] achieve expert-level performance on sleep staging. However, these supervised approaches need large amounts of labeled data and focus mainly on single-modality inputs (usually EEG alone), limiting their ability to capture the multimodal interactions that characterize sleep disorders.

Foundational models learn general-purpose representations from massive unlabeled datasets that transfer to specialized tasks with minimal fine-tuning. SleepFM[28], a multimodal foundation model trained via self-supervised contrastive learning on over 14,000 participants and 100,000+ hours of PSG recordings, is one such approach. By encouraging temporal consistency across EEG, ECG, and respiratory signals, SleepFM learns robust patterns without requiring manual annotations.

Research Question: Can pre-trained foundation model representations(SleepFM) generalize better than statistical features for sleep disorder diagnosis in small, severely imbalanced clinical cohorts, and which algorithmic strategies effectively mitigate data scarcity in this imbalanced cohort?

We test this using the Cyclic Alternating Pattern (CAP) Sleep Database (Section 3), a challenging real-world dataset with 82 full-night recordings across 5 disorder classes with severe class imbalance (majority class 4.4× larger than smallest class). We compare SleepFM embeddings against classical machine learning baselines with statistical features (Section 4), testing class imbalance strategies including adaptive oversampling, class weighting, and focal loss.

Our contributions: (1) empirical comparison between foundation model embeddings and statistical baselines for disorder classification with extreme data scarcity; (2) evaluation of imbalance handling across multiple classifiers; (3) reusable preprocessing pipeline for adapting heterogeneous clinical PSG datasets to foundation model inputs despite channel mismatches; (4) directions for future explainability and clinical validation research.

2 Literature Review

To contextualize our approach, we first review recent advances in deep learning for sleep analysis, foundation models for physiological signals, class imbalance handling strategies, and explainability methods relevant to clinical decision support.

Deep learning has transformed sleep analysis by enabling automatic feature learning from raw physiological signals. U-Sleep[22] demonstrated that convolutional neural networks with U-Net-style encoder-decoder architectures could achieve expert-level sleep staging performance while maintaining computational efficiency. SleepTransformer[23] extended this work by incorporating attention mechanisms to model long-range temporal dependencies, showing improved performance on minority sleep stages.

However, these supervised approaches share common limitations: they require large labeled datasets, focus primarily on sleep staging rather than disorder classification, and typically use single-modality inputs (mainly EEG). This limits clinical usefulness, as sleep disorders often show up through complex patterns across multiple physiological signals.

Foundation models address these limitations through self-supervised learning on massive unlabeled datasets. SleepFM[28] applies multimodal contrastive learning to PSG data, learning representations by maximizing agreement between temporally aligned views while discriminating between different time windows and subjects. This approach offers three advantages: (1) learning from unlabeled data reduces annotation costs; (2) multimodal training captures interactions across EEG, ECG, and respiratory signals; (3) general-purpose representations transfer to various downstream tasks, including disorder classification and disease risk prediction[29].

Class imbalance is common in medical datasets where rare diseases are underrepresented. Standard accuracy metrics become misleading when the majority class prediction achieves high scores without learning meaningful minority class patterns. The imbalanced learning literature[9] identifies three main strategies: (1) data-level methods that resample training distributions (oversampling minorities, undersampling majorities); (2) algorithm-level methods that modify learning objectives (class weighting, focal loss[19]); and (3) ensemble methods combining multiple resampled models. For medical applications, precision-recall metrics have been recommended as more informative than ROC curves[26], though systematic comparisons for foundation model embeddings remain limited.

Explainability research for sleep analysis has explored both intrinsic and post-hoc approaches. Attention mechanisms embedded in model architectures are meant to show important input regions, though studies question whether attention weights reliably indicate feature importance[2]. Post-hoc methods like Integrated Gradients[16] and Grad-CAM[30] have successfully revealed clinically meaningful patterns (sleep spindles, K-complexes) that align with expert scoring criteria, forming a basis for future interpretability research.

3 Dataset: Cyclic Alternating Pattern (CAP) Sleep Database

Having established the theoretical foundation, we now describe the dataset used to evaluate foundation model performance under realistic clinical constraints. The CAP Sleep Database presents several challenges that make it well-suited for testing generalization: severe class imbalance, heterogeneous recording configurations, and limited sample size.

3.1 Structure and Modalities

The CAP Sleep Database is an open-access collection of 108 full-night polysomnography recordings from the Sleep Disorders Center, Ospedale Maggiore in Parma, Italy, distributed via PhysioNet[11]. Originally created to provide expert-annotated examples of Cyclic Alternating Pattern (CAP), an NREM microstructure marker of sleep instability, the dataset includes full sleep stage scoring and diagnostic labels.

Each recording is stored in European Data Format (.edf) with accompanying annotation files (.txt) containing sleep stage transitions (scored according to Rechtschaffen & Kales criteria[20]), CAP phase-A events (subtypes A1/A2/A3), timestamps, and body position when available. The database contains 16 healthy control subjects and 92 pathological recordings distributed across seven sleep disorders: Nocturnal Frontal Lobe Epilepsy (NFLE, n=40), REM Behavior Disorder (RBD, n=22), Periodic Leg Movement Disorder (PLMD, n=10), Insomnia (n=9), Narcolepsy (n=5), Sleep-Disordered Breathing (SDB, n=4), and Bruxism (n=2).

All recordings include core PSG modalities (EEG, EOG, EMG, ECG), though specific channel configurations vary across subjects. EEG channels include frontal, central, and occipital derivations based on the 10-20 system[14], with both referential (e.g., C3-M2) and bipolar (e.g., F4-C4) montages present. Respiratory signals (nasal airflow, thoracic/abdominal effort, oxygen saturation SaO_2, photoplethysmography) are particularly sparse, with only 78.7% of recordings containing SaO_2 measurements.

3.2 Cohort Composition

Gender distribution is approximately 60% male and 40% female, with strong disorder-specific skew (RBD and SDB mainly male, Insomnia more female). Subject ages range from 14 to 82 years, with median age varying substantially by disorder group. This demographic heterogeneity reflects real-world clinical populations but introduces methodological challenges for balanced sampling.

3.3 Compatibility with SleepFM: Preprocessing

For this project, we filtered the dataset to include only recordings with all channels required for SleepFM, resulting in a final cohort of 82 subjects. The primary bottleneck was SaO_2 availability, reducing sample size from 108 to 82. To enable stratified cross-validation, we excluded disorders with fewer than 4 subjects (Narcolepsy, SDB, Bruxism), yielding our final 5-class classification task: Control (n=16), NFLE (n=40), RBD (n=22), PLMD (n=10), and Insomnia (n=9).

Preprocessing steps harmonize heterogeneous channel configurations for SleepFM compatibility. All signals are resampled to 256 Hz, segmented into 30-second epochs, and mapped to SleepFM's expected 9-channel structure (3 respiratory, 5 brain activity, 1 ECG). Since CAP recordings use different channel names and montages, we perform physiologically justified substitutions detailed in Section 4.2. Subject-level splits prevent data leakage, and stratification maintains representative disorder distributions in train-validation-test folds.

4 Methodology

With the dataset and preprocessing constraints established, we now describe our experimental approach. We compare two feature extraction strategies, frozen SleepFM embeddings versus handcrafted statistical features, and systematically evaluate class imbalance mitigation techniques across multiple classifier architectures.

4.1 Model Approach: Downstream Classification With SleepFM Embeddings

We leverage pre-trained SleepFM as a fixed feature extractor for downstream sleep disorder classification. Instead of fine-tuning encoder weights, we freeze all layers and extract multimodal embeddings that serve as input to lightweight classifiers.

Embedding Extraction Each 30-second epoch passes through frozen SleepFM encoders for EEG/EOG/EMG (BAS), ECG, and respiratory modalities. Each encoder produces a modality-specific embedding vector: $z_{BAS}, z_{ECG}, z_{Resp} \in \mathbb{R}^{512}$. These embeddings are L2-normalized and concatenated through SleepFM's multimodal fusion head to form a single joint latent feature: $z_{joint} \in \mathbb{R}^{1536}$.

Subject-Level Aggregation CAP disorder labels are assigned at the subject level, requiring the aggregation of variable-length epoch sequences. For each patient P with T epochs, we compute global statistics:

$$\mu_P = \frac{1}{T}\sum_{t=1}^{T} z_t \tag{1}$$

$$\sigma_P = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(z_t - \mu_P)^2} \tag{2}$$

Concatenating $[\mu_P, \sigma_P]$ produces a 3,072-dimensional subject representation that captures both average physiological state and the variability at night.

Downstream Classifier We evaluate three lightweight classifier architectures trained on aggregated embeddings:

- **Logistic Regression:** Linear model with LASSO (L1) regularization (C=0.1) for sparsity
- **Random Forest:** Ensemble of 400 trees with balanced class weights
- **K-Nearest Neighbors:** Distance-based classifier (k=5) with uniform and distance-weighted voting

4.2 Choice of Channels

SleepFM expects a fixed 9-channel input tensor: 3 respiratory (Chest, SaO_2, Abdomen), 5 brain activity signals (C3-M2, C4-M1, O1-M2, O2-M1, E1-M2), and 1 ECG channel. CAP recordings lack exact matches for several expected channels, requiring physiologically justified substitutions.

4.2.1 Respiratory Modality

Chosen CAP channels: [PLETH, SaO_2, PLETH]

Only 32 of 82 CAP recordings contain chest belt inductance, while 85 include photoplethysmography (PPG). We substitute PPG for chest and abdominal belts based on established physiological equivalence.[17] PPG captures cardiovascular dynamics modulated by respiration through respiratory-induced intensity variation (RIIV)[31]. During obstructive breathing events, intrathoracic pressure changes cause characteristic PPG pulse amplitude fluctuations. Research demonstrates a strong correlation (r=0.8) between esophageal pressure (gold standard for respiratory effort) and PPG-derived effort signals[17], validating PPG as a scientifically defensible proxy. Recent work has further demonstrated PPG's use for interpretable sleep detection using feature-based machine learning[21]. Furthermore, SleepFM was pre-trained on 7 heterogeneous respiratory channels, including pulse readings[28], suggesting the model learned to extract respiratory information from multiple signal types.

4.2.2 Brain Activity Signal Modality

Chosen CAP channels: F4-C4, C4-A1, C4-P4, P4-O2, ROC-LOC

Only 13 CAP recordings contain the expected C3-M2 derivation, necessitating substitution with bipolar derivations and the alternative hemisphere. Recent quantitative analysis found minimal interhemispheric differences in sleep EEG, with median amplitude differences $< 3\,\mu$V and highly similar spectral content across all sleep stages[27]. Both hemispheres contain equivalent information for sleep staging, supporting our use of right-hemisphere derivations.

While sleep phenomena show regional variation (occipital alpha for wakefulness, frontal slow waves for N3), studies comparing single-channel staging found frontal derivations perform as well as or better than central/occipital leads[10, 24]. The AASM scoring manual[1] permits either hemisphere, and empirical validation shows negligible scoring accuracy differences between F4/C4/O2 and F3/C3/O1 montages[7].

For EOG, we substitute bipolar ROC-LOC (right-left outer canthus) for the expected E1-M2 referential montage. This horizontal derivation strongly captures rapid eye movements during REM sleep, as eyes move out-of-phase, producing large potential differences. Although amplitude and polarity differ from mastoid-referenced EOG, the underlying physiological signal remains equivalent.

4.3 Extraction/Preprocessing

Two parallel preprocessing pipelines have been implemented: statistical features for classical ML baselines, and SleepFM embeddings for the foundation model approach. Both share core signal processing to ensure fair comparison.

4.3.1 Shared Signal Processing

Raw PSG data undergoes standardized preprocessing:

- **Resampling:** All signals resampled to 256 Hz using MNE-Python
- **Epoch Segmentation:** Continuous recordings segmented into non-overlapping 30-second epochs (7,680 samples per epoch)
- **Label Mapping:** Sleep stage annotations converted from R&K to AASM standard (Wake, N1, N2, N3, REM), excluding Movement/Unscored/CAP subtype epochs

- **SaO$_2$ Artifact Cleaning:** Gradient-based detection of sensor artifacts (drops $\leq -10\%$ with 5-second recovery), repaired via linear interpolation
- **Amplitude Scaling:** EEG/EOG/ECG scaled to microvolts, respiratory signals retain native units

4.3.2 Baseline Pipeline: Statistical Features

For each 30-second epoch, we extract time-domain and frequency-domain features per channel:

Time-domain features:

- Standard deviation: $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$
- Root mean square: $\text{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$
- Line length: $\text{LL} = \sum_{i=1}^{N-1} |x_{i+1} - x_i|$

Frequency-domain features: Relative spectral power in standard EEG bands (Delta: 0.5-4 Hz, Theta: 4-8 Hz, Alpha: 8-13 Hz, Sigma: 12-16 Hz, Beta: 16-30 Hz) computed via Welch's method with 4-second Hamming windows.

Features are aggregated (mean, standard deviation) across all valid epochs in a recording to produce a single feature vector per subject for night-level disorder classification.

4.3.3 SleepFM Pipeline

Epochs (9 channels × 7,680 samples) are passed through frozen SleepFM encoders. Output embeddings are L2-normalized, concatenated, and aggregated via global statistics (μ_P, σ_P) to form 3,072-dimensional subject representations as described in Section 4.1.

4.4 Baseline Model

Classical machine learning baselines use statistical features to establish a performance comparison. Features are extracted from eight channels: PLETH, SaO$_2$, F4-C4, C4-A1, C4-P4, P4-O2, ROC-LOC, ECG1-ECG2. Each 30-second epoch yields time-domain (std, RMS, line length) and frequency-domain (spectral power per EEG band) features, aggregated to subject-level for disorder classification.

We evaluate three classical architectures:

- **Random Forest:** Ensemble of 400 decision trees with unrestricted depth
- **K-Nearest Neighbors:** Distance-based classifier (k=7) with uniform and distance-weighted voting
- **Multi-Layer Perceptron:** Fully-connected network with two hidden layers (128 neurons each), ReLU activation, Adam optimizer

All models are evaluated using 50-seed, 4-fold stratified cross-validation (200 total train-test evaluations) to ensure reliable performance estimates despite a small sample size. Data leakage is prevented through subject-level splits and fold-specific preprocessing (imputation, scaling fit only on training data).

4.5 SleepFM

4.5.1 Sleep Staging Classification with SleepFM

To validate our channel substitution approach, we first evaluated SleepFM on epoch-level sleep staging using frozen encoders + logistic regression classifier with balanced class weights. The model successfully generalized to CAP despite distribution shift, achieving competitive accuracy for N2/N3 stages and reasonable REM detection, confirming that channel substitutions preserve sleep-relevant features.

4.5.2 Night-level Sleep Disorder Classification

Transitioning from epoch-level staging to night-level disorder classification required three modifications: (1) metadata preservation during embedding generation to link embeddings with patient identifiers; (2) temporal reconstruction grouping epochs by patient and sorting chronologically; (3) disorder label derivation from patient identifier prefixes using a mapping directory (e.g., nfle→1, rbd→4).

Subject-level aggregation via global statistics (μ_P, σ_P) produces fixed-dimensional representations despite variable recording lengths, enabling standard classifier training.

5 Addressing Class Imbalance for Sleep Disorder Classification

Given the severe class imbalance in our dataset (majority class 4.4× larger than smallest), we now systematically evaluate strategies to mitigate this challenge. This section describes the theoretical rationale for each approach and our experimental protocol for fair comparison.

5.1 General Problem Description

Class imbalance affects both sleep staging and downstream disorder classification in clinical datasets. The CAP cohort exhibits severe imbalance (NFLE: 40 samples, Insomnia: 9 samples), biasing classifiers toward majority classes and degrading minority class performance, which is a phenomenon extensively studied in imbalanced learning literature[12, 9].

As an initial stabilization step, we removed the two classes with fewer than 5 subjects (Narcolepsy, SDB). From a theoretical perspective, extreme class rarity provides insufficient variability for stable decision boundaries, increasing noise and destabilizing classifiers in small datasets[9].

5.2 Theoretical Techniques Tried on Our Baseline Models

After rare class removal, we systematically evaluated five imbalance strategies:

(A) None/Baseline: Standard training without imbalance handling

(B) Stratified Sampling: Maintain class proportions in train-test splits

(C) Class-Weighted Loss: Assign loss weights inversely proportional to class frequency: $w_c = \frac{N}{k \cdot n_c}$ This technique is commonly recommended as a baseline method for imbalanced classification problems and is supported by both theoretical and empirical studies[5, 18].

(D) Random Oversampling: Duplicate minority class samples until balanced distribution. Techniques such as random oversampling or more advanced synthetic methods like SMOTE have been widely studied and shown to improve minority-class recall in certain settings[6, 18]. However, prior work emphasizes that when minority classes contain very few samples, oversampling may increase the risk of overfitting by repeatedly exposing the model to highly similar examples[3, 9]. To mitigate this, we evaluate two noise injection variants:

- *Constant noise:* Add Gaussian noise $\mathcal{N}(0, \epsilon)$ uniformly across all features
- *Adaptive noise:* Scale noise by feature standard deviation: $\mathcal{N}(0, \epsilon \cdot \sigma_{\text{feature}})$

(E) Focal Loss: Down-weight easy examples: $\mathcal{L}_{\text{focal}} = -\alpha_t(1 - p_t)^\gamma \log(p_t)$

For KNN, traditional loss-based techniques cannot be applied. We evaluate distance-weighted voting where closer neighbors contribute more strongly, providing partial sensitivity to local class distributions[8].

5.3 Experiments on Our Baseline

All experiments were repeated over 50 random seeds with 4-fold stratified cross-validation to obtain stable performance estimates despite the small dataset size (82 subjects). Averaging over multiple seeds reduces impact of favorable/unfavorable splits.

5.3.1 Rare-Class Removal

Comparing performance before/after filtering on our old code (single train-test split without stratified CV):

Random Forest: Accuracy increased from 0.62 to 0.70, while Macro-F1 improved from 0.25 to 0.44, indicating that the removed classes contributed primarily to majority-class bias

KNN: Accuracy is stable at 0.62, but the Macro-F1 increased from 0.32 to 0.43

These results motivated our decision to work with the filtered 5-class problem for all subsequent experiments with our new codes.

5.3.2 Results on Random Forest with Class Imbalance Mitigation

The baseline without imbalance correction achieved Macro-F1=0.35, AUROC=0.79. Class-weighted loss provided marginal improvement (Macro-F1=0.36), while focal weighting slightly decreased performance (Macro-F1=0.34).

Oversampling with adaptive noise consistently and slightly improved Macro-F1 as the noise level increased (see Appendix Figure 9), with optimal performance at 10% adaptive noise (Macro-F1=0.44, AUROC=0.80). Constant noise degraded performance, distorting relationships between small-scale and large-scale features. Adaptive scaling preserves relative feature relationships while introducing controlled variability.

5.3.3 Experiments on the KNN Model

Uniform weighting achieved Macro-F1=0.25, AUROC=0.67. Distance-weighted voting greatly improved performance (Macro-F1=0.32, AUROC=0.70), confirming that prioritizing closer neighbors partially alleviates majority-class dominance.

Oversampling with adaptive noise further improved performance (see Appendix Figure 10), with optimal configuration at 12% adaptive noise + distance weighting (Macro-F1=0.38, AUROC=0.69). However, KNN remains fundamentally limited by local majority voting, constraining its ability to fully address severe class imbalance.

5.3.4 Experiments on the MLP's Model

Baseline MLP achieved Macro-F1=0.44, AUROC=0.73. Class-weighted cross-entropy reduced Macro-F1 to 0.28 (AUROC=0.74). This suggests that strong reweighting may have made optimization less stable or increased sensitivity to scarce minority-class samples, which can be problematic in small datasets. Focal loss improved both metrics (Macro-F1=0.45, AUROC=0.74), confirming its ability to reduce easy-example influence.

Oversampling with adaptive noise achieved the best performance at 10% noise level (Macro-F1=0.46, AUROC=0.74), representing a clear improvement over baseline while preserving discriminative power.

5.3.5 Comparison of Imbalance Mitigation Strategies

Data-level augmentation (adaptive oversampling) consistently outperformed loss-based reweighting across all architectures. For Random Forest and MLP, oversampling with feature-scaled noise yielded the largest Macro-F1 improvements while preserving AUROC. However, it should be noted that adding adaptive noise to oversampling provided only marginal additional gains compared to simple oversampling, suggesting that the main benefit comes from rebalancing class frequencies rather than from the injected variability itself. Class-weighted and focal losses provided limited or inconsistent gains, particularly in data-scarce settings. For KNN, distance weighting improved performance, with further gains when combined with adaptive oversampling. Adaptive noise consistently outperformed constant noise, highlighting the importance of scaling perturbations to feature distributions.

5.4 SleepFM

Foundation model embeddings exhibited different imbalance sensitivity compared to statistical features. We evaluated five scenarios: (A) None/Baseline, (B) Class-weighted loss, (C) Oversampling with adaptive noise, (E) Focal weighting.

Key findings:

Random Forest baseline achieves strong performance without mitigation: Macro-F1=0.41, AUROC=0.87, AUPRC=0.71, substantially outperforming all baseline statistical feature models. Imbalance mitigation provided marginal or negative returns, with class weighting and focal loss degrading performance. This suggests pre-trained SleepFM embeddings already provide enough separability that heavy rebalancing prioritizes noise over signal.

KNN benefits from oversampling: Distance-based classifiers still gained from data augmentation (Macro-F1: 0.36→0.46 with oversampling), consistent with baseline feature results.

Linear models underperform: Logistic regression achieved only Macro-F1=0.28 baseline, improving to 0.34 with oversampling, indicating the 3,072-dimensional embedding space contains non-linear decision boundaries better captured by ensemble methods.

6 Results

Having described our methodology and experimental design (Section 4), we now present the empirical results. We first report baseline model performance with statistical features, then compare against SleepFM embeddings, and finally assess the effectiveness of different imbalance mitigation strategies.

6.1 Evaluation Metrics

Accuracy measures the proportion of correctly classified samples over all samples. While it is easy to interpret, accuracy can be misleading in imbalanced datasets, as high values may be obtained by favoring majority classes.

Macro F1-score is computed by first calculating the F1-score independently for each class and then averaging them. This metric assigns equal importance to all classes, making it more appropriate for multi-class sleep disorder classification where rare disorders should be evaluated fairly.

AUROC (Area Under the ROC Curve) evaluates the model's ability to distinguish between classes by analyzing the trade-off between the true positive rate and false positive rate across different decision thresholds. In the multi-class setting, AUROC is computed using a one-vs-rest strategy and reflects overall discriminative performance.

AUPRC (Area Under the Precision–Recall Curve) summarizes the trade-off between precision and recall across thresholds. It is particularly informative for imbalanced datasets, as it emphasizes correct identification of minority classes, which is crucial in clinical sleep disorder detection.

6.2 Baseline Models: Statistical Features

6.2.1 Effect of Rare Class Removal

Comparing performance before/after filtering on our old code (single train-test split without stratified CV):

Random Forest: Accuracy increased from 0.62 to 0.70, while Macro-F1 improved from 0.25 to 0.44, indicating that the removed classes contributed primarily to majority-class bias

KNN: Accuracy is stable at 0.62, but the Macro-F1 increased from 0.32 to 0.43

These results motivated our decision to work with the filtered 5-class problem for all subsequent experiments with our new codes.

6.2.2 K-Nearest Neighbors Performance

Figure 1 shows that the single stratified split consistently achieves higher accuracy (up to =0.65 for k=5–11) than 4-fold cross-validation, whose mean accuracy peaks at around

0.56. The wide ±1 SD region indicates substantial fold-to-fold variability, highlighting sensitivity to the chosen data split and the optimistic nature of single-split evaluation.

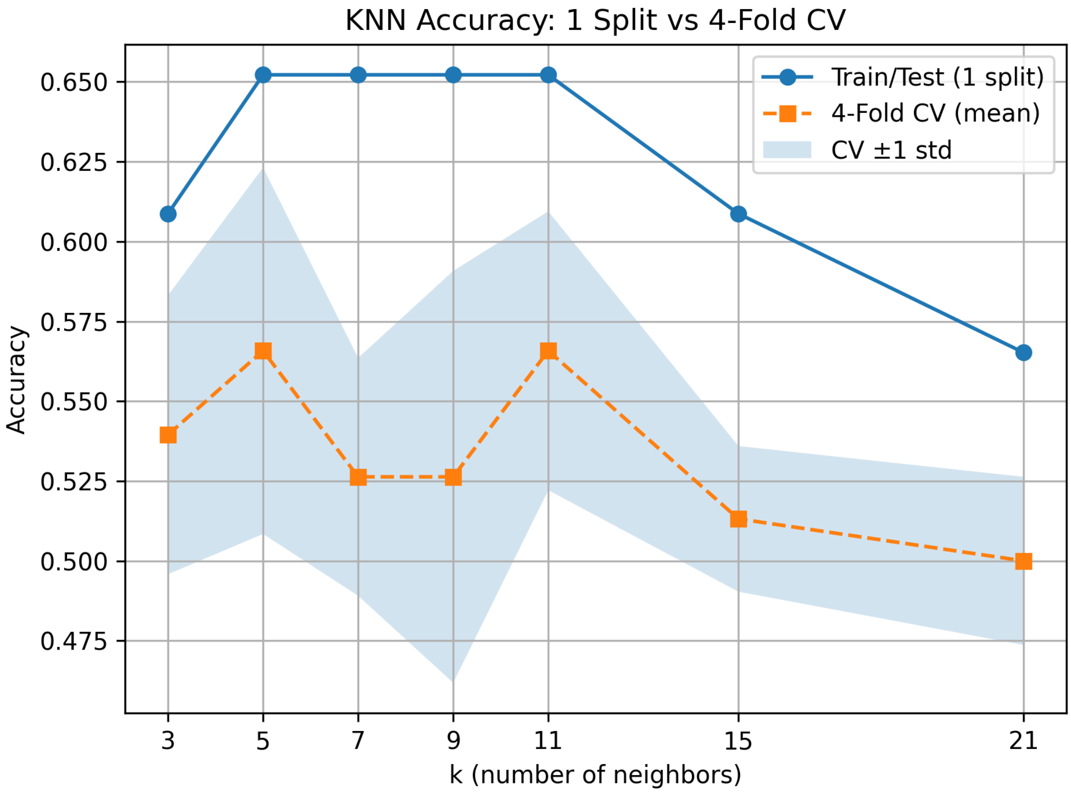


Figure 1: KNN accuracy comparison between single stratified split and 4-fold CV. CV provides more reliable estimates with explicit performance variability quantification.

Macro-F1 results (Figure 2) show that the single stratified split consistently outperforms 4-fold cross-validation, peaking at approximately 0.43 for k=3–7, while CV achieves a lower maximum Macro-F1 of about 0.31 at k=5. Performance declines steadily for larger k values, indicating increased bias toward majority classes and reduced minority-class recall.

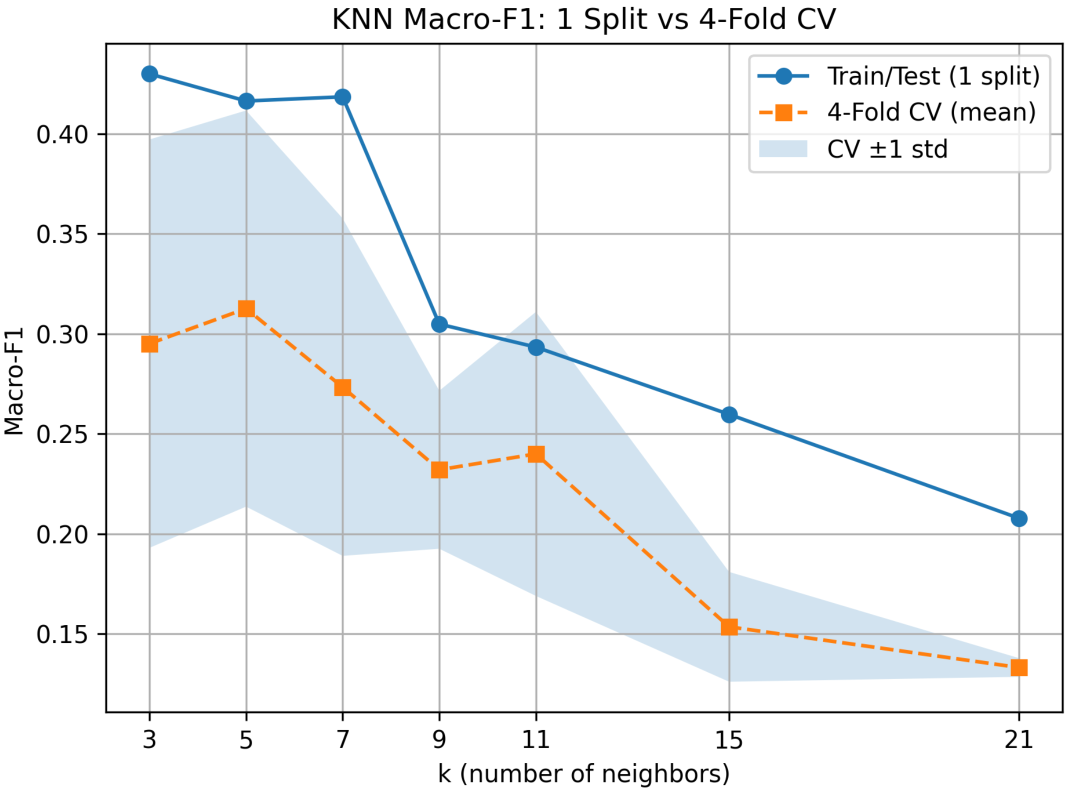
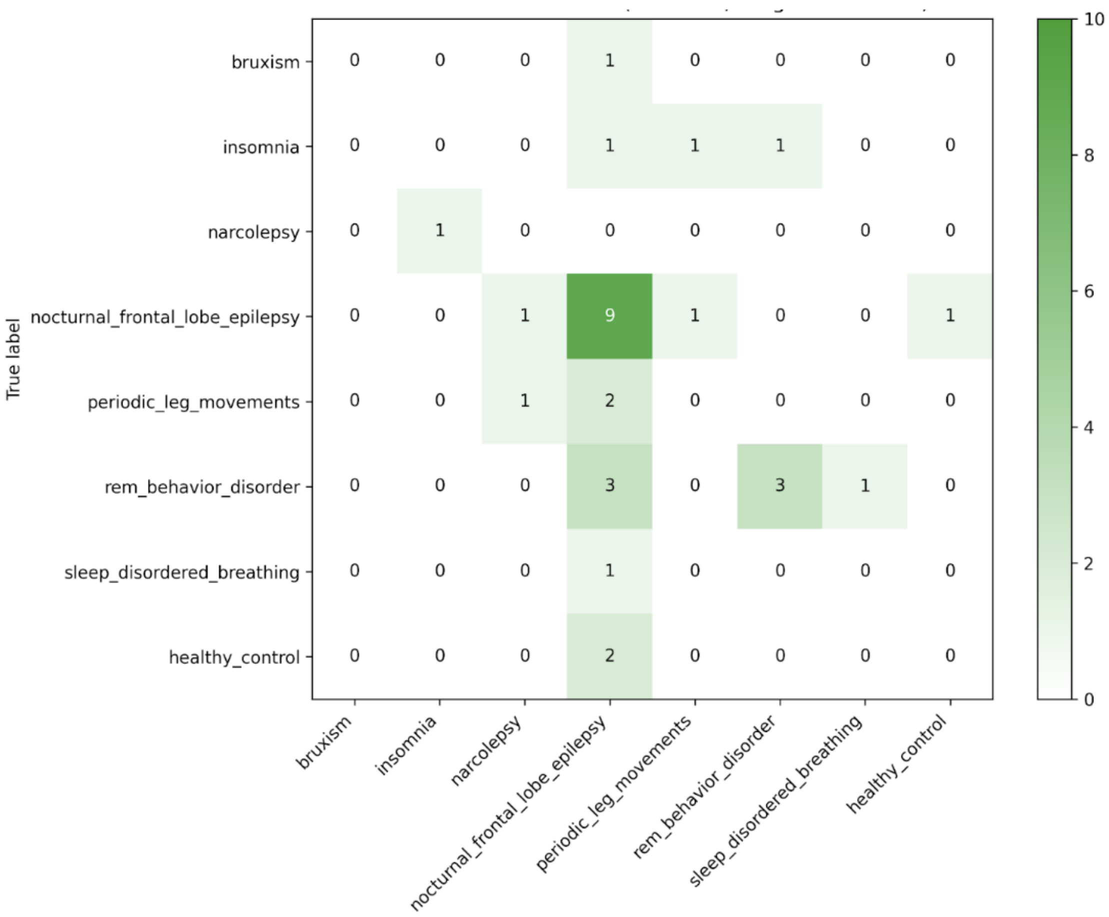


Figure 2: KNN Macro-F1 comparison. CV yields lower but more stable performance, with k=3 selected in four of five folds.

Confusion matrices (Figures 3, 4) reveal consistent misclassification patterns: NFLE most accurately classified, while RBD and PLMD are frequently confused with other disorders.



	bruxism	insomnia	narcolepsy	nocturnal_frontal_lobe_epilepsy	periodic_leg_movements	rem_behavior_disorder	sleep_disordered_breathing	healthy_control
bruxism	0	0	0	1	0	0	0	0
insomnia	0	0	0	1	1	1	0	0
narcolepsy	0	1	0	0	0	0	0	0
nocturnal_frontal_lobe_epilepsy	0	0	1	9	1	0	0	1
periodic_leg_movements	0	0	1	2	0	0	0	0
rem_behavior_disorder	0	0	0	3	0	3	1	0
sleep_disordered_breathing	0	0	0	1	0	0	0	0
healthy_control	0	0	0	2	0	0	0	0

Figure 3: KNN confusion matrix (single split, k=7, distance-weighted). Strong NFLE performance, poor minority class discrimination.

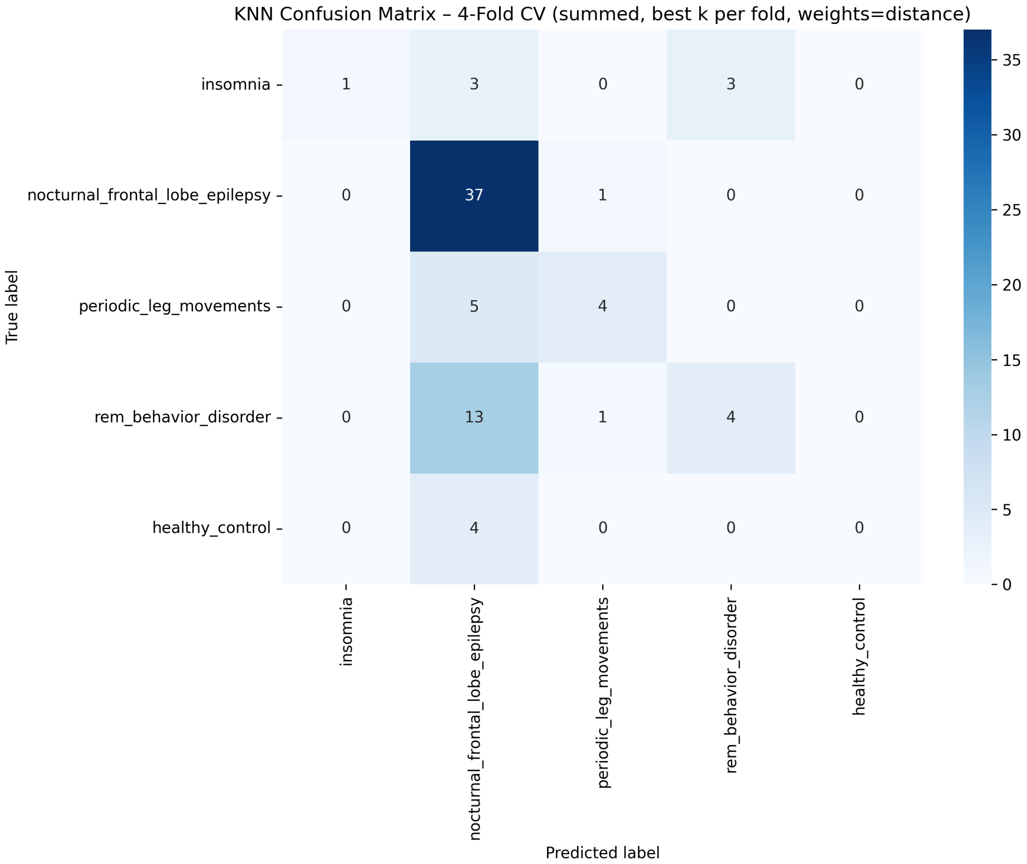


Figure 4: KNN confusion matrix (4-fold CV, summed across folds). Consistent pattern: NFLE well-classified, RBD/PLMD frequently misclassified.

6.2.3 Random Forest Performance

Random Forest on a single split achieved accuracy=0.60, Macro-F1=0.26 (Figure 5). Low Macro-F1 despite moderate accuracy indicates strong majority-class bias.

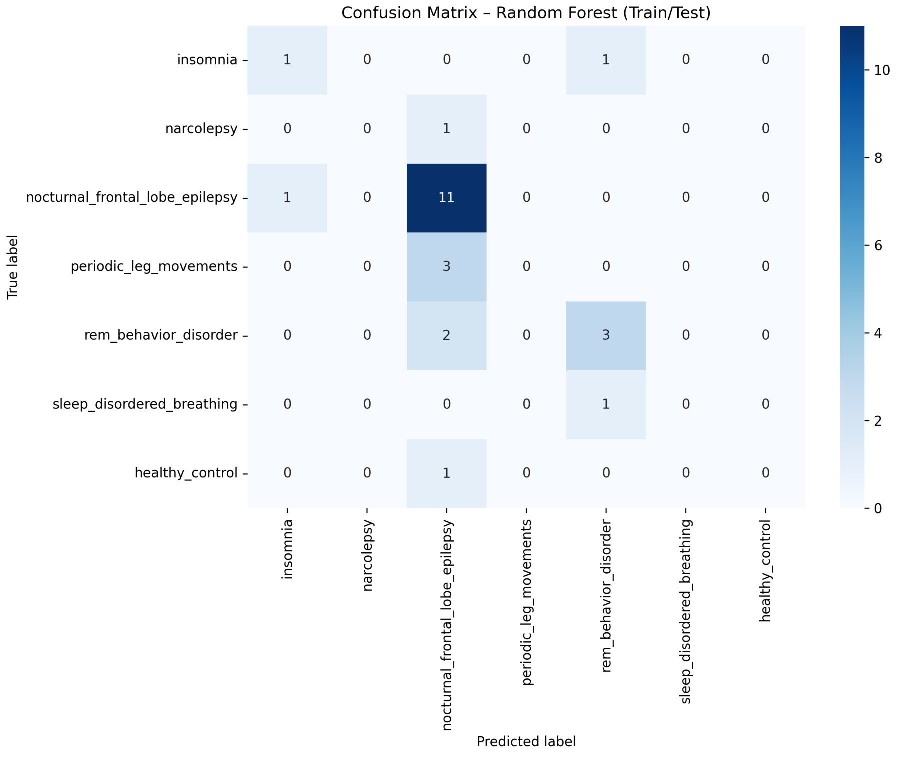


Figure 5: Random Forest confusion matrix (single split). Correct predictions concentrated on NFLE; minority classes were frequently misclassified.

Cross-validation improved performance (accuracy=0.78, Macro-F1=0.35), but a modest Macro-F1 gain reveals persistent class imbalance challenges (Figure 6). The normalized confusion matrix shows NFLE dominates predictions, with Control, PLMD, and Insomnia often misclassified as NFLE.

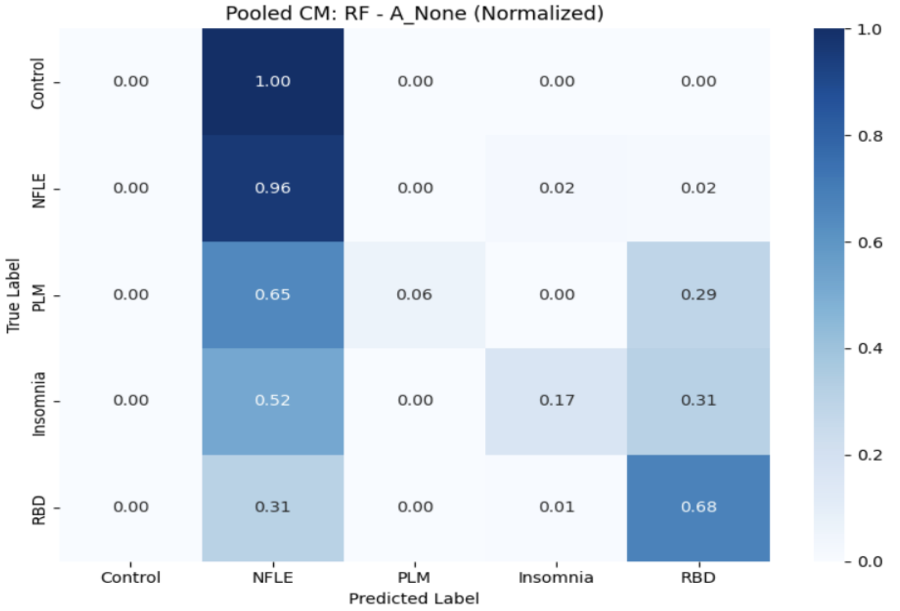


Figure 6: Random Forest normalized confusion matrix (4-fold CV). Persistent NFLE bias despite cross-validation. RBD shows comparatively better detection.

6.2.4 Class Imbalance Mitigation: Baseline Models

Table 1 summarizes results from evaluation of strategies (50 seeds, 4-fold CV, 200 evaluations per configuration).

Model	**Strategy**	**Macro-F1**	**AUROC**
RF	Baseline	0.35	0.79
RF	Class-weighted	0.36	0.80
RF	Focal	0.34	0.80
RF	**Oversample + Adaptive (10%)**	**0.44**	0.80
KNN	Uniform	0.25	0.67
KNN	Distance-weighted	0.32	0.70
KNN	**Oversample + Distance + Adaptive (12%)**	**0.38**	0.69
MLP	Cross-entropy	0.44	0.73
MLP	Weighted CE	0.28	0.74
MLP	Focal loss	0.45	0.74
MLP	**Oversample + Adaptive (10%)**	**0.46**	0.74

Table 1: Imbalance mitigation comparison for baseline models (50 seeds, 4-fold CV).

Key findings: Data-level methods outperform loss-based reweighting. Oversampling with adaptive noise achieved the largest Macro-F1 improvements across all architectures while preserving AUROC. Loss-based methods provided limited or inconsistent gains, particularly in high-dimensional spaces with few samples. Distance weighting improved KNN, with further gains when combined with adaptive oversampling.

6.3 SleepFM Foundation Model Results

6.3.1 Sleep Staging Validation

Epoch-level sleep staging using frozen SleepFM + balanced logistic regression successfully generalized to CAP despite distribution shift, achieving 0.7 accuracy and very competitive accuracy for N2/N3 stages and reasonable REM detection (see Appendix Figure 12). This validates that channel substitutions preserve sleep-relevant features.

6.3.2 Disorder Classification Performance

Table 2 presents disorder classification results for lightweight classifiers trained on aggregated SleepFM embeddings.

Model	**Strategy**	**Macro-F1**	**AUROC**	**AUPRC**
Logistic Regression	None (baseline)	0.28	0.75	0.53
Logistic Regression	**Oversampling**	**0.34**	0.75	0.51
Random Forest	**None (baseline)**	**0.41**	**0.87**	**0.71**
Random Forest	Class-weighted loss	0.34	0.82	0.67
Random Forest	Oversampling	0.40	0.86	0.69
Random Forest	Focal weighting	0.38	0.80	0.64
KNN	None (baseline)	0.36	0.76	0.56
KNN	**Oversampling**	**0.46**	0.80	0.63

Table 2: Disorder classification using SleepFM embeddings (50 seeds, 4-fold CV). Random Forest baseline achieves the best overall performance without requiring imbalance mitigation.

Key findings:

Foundation model embeddings enable strong baseline performance. Random Forest on raw SleepFM embeddings (no imbalance handling) achieved Macro-F1=0.41, AUROC=0.87, AUPRC=0.71, substantially outperforming all baseline statistical feature models and rivaling their best imbalance-mitigated configurations.

Imbalance mitigation provides marginal or negative returns. Unlike baseline features, foundation model embeddings showed minimal benefit from rebalancing techniques. Class weighting and focal loss degraded Random Forest performance, suggesting pre-trained SleepFM embeddings already provide enough separability that heavy rebalancing prioritizes noise over signal.

KNN benefits from oversampling. Distance-based classifiers still gained from data augmentation (Macro-F1: 0.36→0.46), consistent with baseline feature results.

6.4 Cross-Method Comparison

Approach	**Features**	**Model**	**Macro-F1**	**AUROC**
Baseline	Handcrafted	MLP + OS (A10%)	0.46	0.74
Baseline	Handcrafted	RF + OS (A10%)	0.44	0.80
SleepFM	**Frozen**	**RF (none)**	**0.41**	**0.87**
SleepFM	Frozen	KNN + OS	0.46	0.80

Table 3: Best-performing configurations across feature representations. SleepFM + RF achieves the highest AUROC without imbalance handling.

Foundation model embeddings provide three advantages: (1) Superior discriminative power (AUROC=0.87 vs. 0.74-0.80 for statistical features); (2) Reduced reliance on data augmentation (strong baseline without oversampling); (3) Simpler training pipeline (no hyperparameter tuning for imbalance mitigation).

However, carefully tuned baseline models with adaptive oversampling can match or exceed foundation model Macro-F1 (0.46 vs. 0.41-0.46), suggesting statistical features retain value when imbalance is properly addressed.

7 Discussion

We now interpret our experimental results (Section 6) in the context of the research question: can pre-trained foundation model embeddings work for sleep disorder diagnosis in small, imbalanced clinical datasets? Our findings reveal both advantages and limitations of this approach, with implications for future clinical deployment.

7.1 Foundation Models vs. Statistical Features

Pre-trained foundation model embeddings provide advantages for disorder classification in small, imbalanced clinical cohorts. Foundation model embeddings enabled Random Forest to achieve AUROC=0.87 without imbalance handling, outperforming all baseline statistical feature configurations. This suggests that self-supervised contrastive pre-training on 100,000+ hours of PSG data learned general-purpose physiological patterns that transfer effectively despite channel substitutions and severe domain shift.

However, the Macro-F1 gap is smaller (0.41 vs. 0.46), with carefully tuned baseline models matching or slightly exceeding SleepFM when combined with oversampling. This indicates handcrafted spectral and temporal features retain discriminative value for minority classes when properly balanced, possibly because they explicitly encode frequency-domain patterns (delta, theta, alpha power) that domain experts know to be clinically relevant.

The key trade-off is development effort: achieving competitive baseline performance required systematic exploration of noise injection strategies, hyperparameter tuning, and model-specific imbalance handling. In contrast, SleepFM provided strong results with minimal tuning, suggesting foundation models can reduce the expertise barrier for clinical ML applications.

7.2 Class Imbalance Mitigation Strategies

Across both feature types, data-level augmentation (adaptive oversampling) consistently outperformed loss-based reweighting. This finding aligns with imbalanced learning literature[9] emphasizing that synthetic sample generation increases minority-class decision boundary stability, whereas loss weighting can amplify noise when few real samples exist.

The critical insight from noise injection experiments is that adaptive noise scaled to feature variance outperforms constant noise across all architectures. Constant noise applied uniform perturbations regardless of feature magnitude, distorting relationships between small-scale and large-scale features. Adaptive scaling preserves relative feature relationships while introducing controlled variability, preventing overfitting to identical duplicated samples. However, the dominant performance gain is still driven by oversampling itself, as the addition of noise only leads to marginal improvements.

For foundation model embeddings specifically, imbalance mitigation provided marginal or negative returns. This suggests SleepFM's latent space already exhibits enough class separability that heavy rebalancing prioritizes noise over actual physiological patterns, a practical advantage for clinical deployment.

7.3 Methodological Considerations

Our strict subject-level splits and fold-specific preprocessing prevented common leakage sources[25, 4]: global statistics computed on the entire dataset, family structure leakage from related recordings, and epoch-level splits allowing the same patient in train and test. The 50-seed, 4-fold protocol yielded 200 independent evaluations, providing reliable performance estimates despite a small sample size.

While our physiological justification for channel substitutions is scientifically grounded, the resulting distribution shift likely degraded absolute performance relative to ideal matched channels. The successful sleep staging validation provides evidence that critical sleep structures remain detectable despite substitutions, but quantifying the performance penalty would require recordings with both SleepFM-expected and CAP-available channels, data we lack.

We deliberately avoided fine-tuning SleepFM encoders due to extreme data scarcity (82 subjects) and risk of catastrophic forgetting. However, this constraint limited our ability to adapt representations to CAP-specific patterns. Recent work on parameter-efficient fine-tuning (LoRA, adapters)[29] suggests partial adaptation may improve performance without full re-training.

7.4 Clinical Implications

AUROC=0.87 achieved by SleepFM + Random Forest suggests automatic disorder screening is feasible in real-world clinical settings with heterogeneous recording equipment and limited labeled data. However, several challenges remain before clinical deployment:

Interpretability: Current predictions lack explanations, limiting clinician trust. Section 8 outlines future XAI research to address this.

Generalization: Our evaluation is limited to a single dataset from one Italian sleep center. Multi-site validation is essential to assess robustness to population differences, recording protocols, and annotation variability.

Minority class performance: Macro-F1=0.41 indicates remaining difficulties with rare disorders (PLMD, Insomnia), likely requiring targeted data collection or disorder-specific architectures.

Clinical workflow integration: Automatic screening complements rather than replaces expert diagnosis. Appropriate use cases include prioritizing cases for expert review, quality control flagging, and educational feedback for trainees.

8 Future Work

While our results establish that foundation models can classify sleep disorders effectively even with limited data, several research directions remain unexplored. We organize future work into four priorities: explainability methods, algorithmic improvements, dataset expansion, and clinical deployment pathways.

8.1 Explainability

While this project established strong classification performance, the decision-making process remains opaque. Future work should prioritize transitioning this framework into an explainable research platform to identify physiological biomarkers and build clinical trust.

8.1.1 Localization of Pathological Patterns

The primary XAI objective is identifying which signal modalities and time windows drive predictions. We recommend two complementary approaches:

Hard attention mechanisms: Unlike standard soft attention, hard attention forces discrete segment selection, producing more stable and medically defensible explanations[15]. Integrating hard attention into the SleepFM fusion head would enable automated modality attribution, revealing whether the model prioritizes EEG vs. ECG vs. respiratory signals for different disorders.

Post-hoc saliency mapping: Signal-space attribution methods validated by domain experts:

- *Compensated Integrated Gradients*[16]: Addresses baseline selection challenges specific to physiological signals, providing reliable electrode-level importance scores
- *1D Grad-CAM*[30]: Highlights high-frequency events (K-complexes, sleep spindles, slow waves) that align with manual scoring criteria, enabling clinical validation of learned features

Explanations should be systematically validated by comparing model-highlighted regions against expert annotations (CAP phase-A events, sleep stage transitions) to assess physiological plausibility.

8.1.2 Discovery of Hidden Digital Biomarkers

Beyond validating known patterns, XAI could uncover subtle non-linear features missed by conventional manual scoring. Foundation models trained on massive datasets may detect early-stage disorder signatures or cross-modal interactions (e.g., ECG-EEG coupling during arousal events) not explicitly annotated[29]. Systematic analysis of high-importance regions across large cohorts could identify novel diagnostic markers for prospective validation.

8.2 Algorithmic and Architectural Improvements

Parameter-efficient fine-tuning: While full fine-tuning risks overfitting on 82 subjects, methods like Low-Rank Adaptation (LoRA) allow targeted encoder adaptation with minimal trainable parameters. Pilot studies should compare frozen vs. partially adapted embeddings.

Temporal modeling: Current aggregation (μ_P, σ_P) discards epoch ordering. Recurrent or Transformer-based sequence models could capture the nocturnal evolution of sleep architecture (e.g., REM density changes in RBD). However, this requires careful regularization given limited samples.

Multi-task learning: Jointly training disorder classification with auxiliary tasks (sleep staging, arousal detection, CAP phase prediction) could improve representation quality through shared inductive biases, similar to recent multi-task SleepFM extensions[29].

8.3 Dataset Expansion and Multi-Site Validation

The CAP database represents a single Italian sleep center's patient population and recording protocols. Robust clinical validation requires:

Cross-dataset evaluation: Testing on independent cohorts (SHHS, MASS, ISRUC) to assess generalization across demographics, equipment, and annotation standards

Targeted data collection: Increasing samples for rare disorders (PLMD, Insomnia) through multi-center collaboration or federated learning approaches

Longitudinal studies: Evaluating disorder progression prediction using repeated PSG recordings from the same patients over months/years

8.4 Clinical Deployment Pathways

Transitioning from research prototype to clinical tool requires:

Uncertainty quantification: Bayesian deep learning or ensemble methods to provide calibrated confidence scores, enabling risk-stratified triage (high-confidence predictions for automated processing, low-confidence for expert review)

Human-in-the-loop workflows: Interactive systems where automatic screening shows cases for prioritized expert attention, with explainability tools supporting diagnostic reasoning

Regulatory validation: Prospective clinical trials comparing automatic+expert vs. expert-only workflows to demonstrate safety, efficacy, and time savings under real-world conditions

9 Conclusion

This project investigated whether pre-trained SleepFM embeddings can enable reliable automatic sleep disorder classification in small, severely imbalanced clinical cohorts. Our full evaluation on 82 CAP database recordings across 5 disorders demonstrated that foundation model embeddings provide advantages over handcrafted statistical features, achieving AUROC=0.87 with Random Forest without requiring explicit class imbalance mitigation.

Three key findings emerged: (1) foundation model embeddings exhibited superior separability in latent space, reducing dependence on heavy rebalancing techniques that dominated baseline feature performance; (2) Data-level augmentation via adaptive oversampling consistently outperformed loss-based reweighting across all architectures, with noise scaling critical to preserving feature relationships; (3) Careful channel substitution enabled SleepFM generalization despite severe distribution shift, though absolute performance likely remains below ideal matched-channel scenarios.

While our best configurations achieved competitive performance (Macro-F1: 0.41-0.46), large challenges remain for clinical deployment: minority class performance needs improvement (particularly PLMD and Insomnia), multi-site validation is essential to confirm generalization, and explainability research is critical for building clinician trust. The roadmap outlined in Section 8, combining hard attention mechanisms, validated saliency mapping, and prospective clinical trials, provides a path toward transparent, trustworthy automatic disorder screening systems.

This work establishes a rigorous baseline for sleep disorder classification under realistic data constraints and shows that modern foundation models can extract meaningful physiological patterns from heterogeneous clinical datasets with minimal labeled data. As pre-training datasets continue to grow and explainability methods mature, automatic sleep analysis has the potential to augment expert diagnosis, accelerate patient triage, and ultimately improve access to high-quality sleep medicine.

10 Appendix

10.1 Baseline Random Forest Confusion Matrices

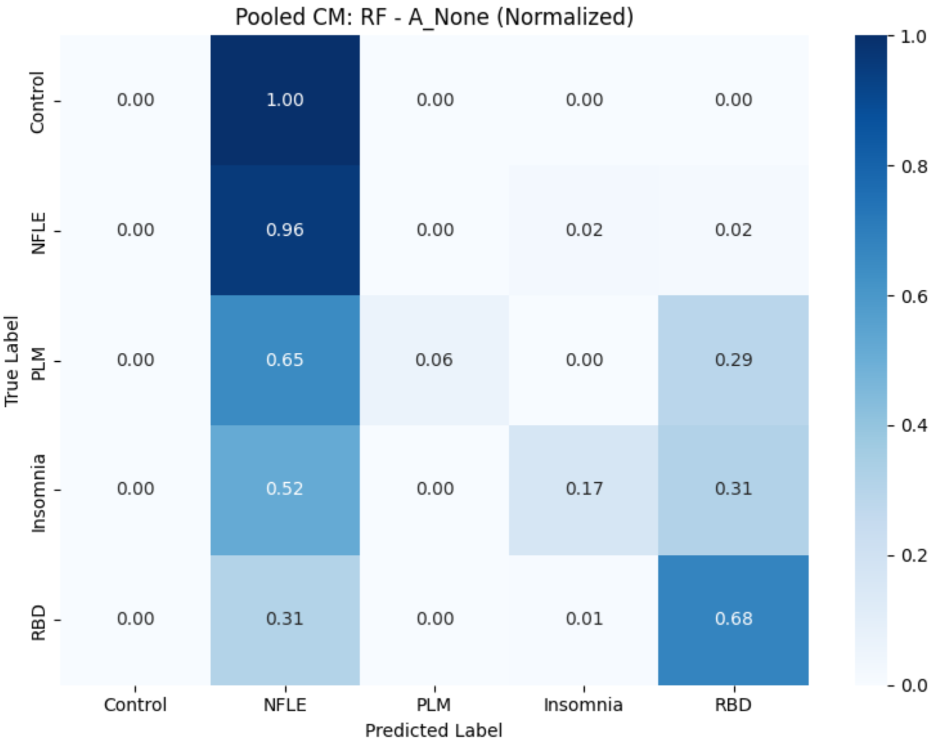


Figure 7: Random Forest baseline confusion matrix (normalized, no imbalance mitigation). Shows strong NFLE classification but poor minority class performance, motivating imbalance handling strategies.

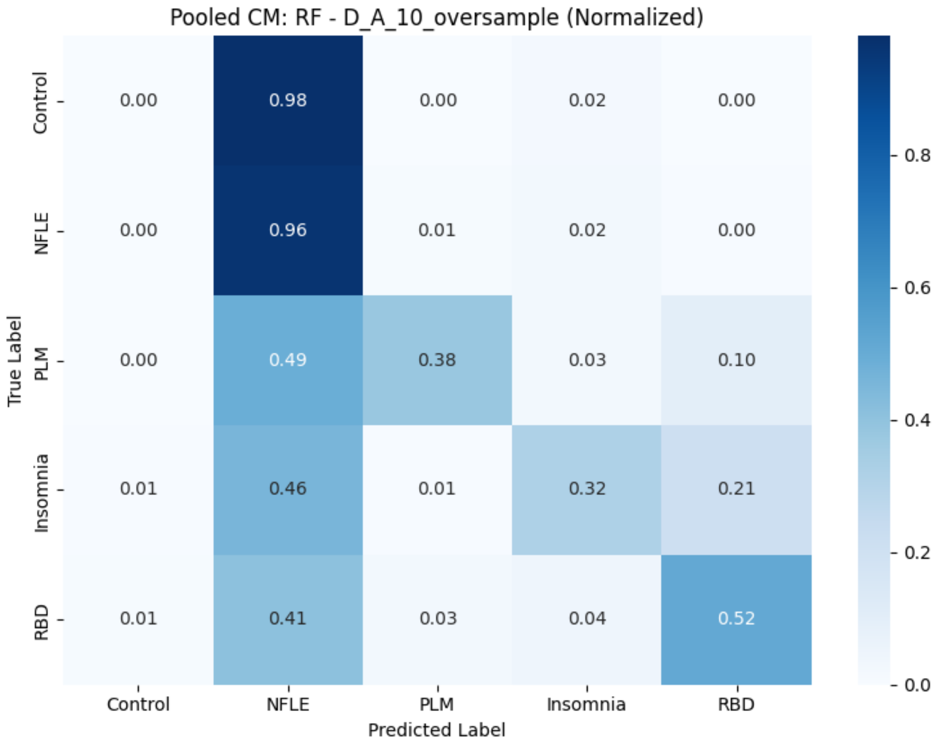


Figure 8: Random Forest confusion matrix with 10% adaptive noise oversampling (normalized). Improved minority class detection compared to baseline, particularly for PLMD and Insomnia.

10.2 Noise Level Ablation Studies

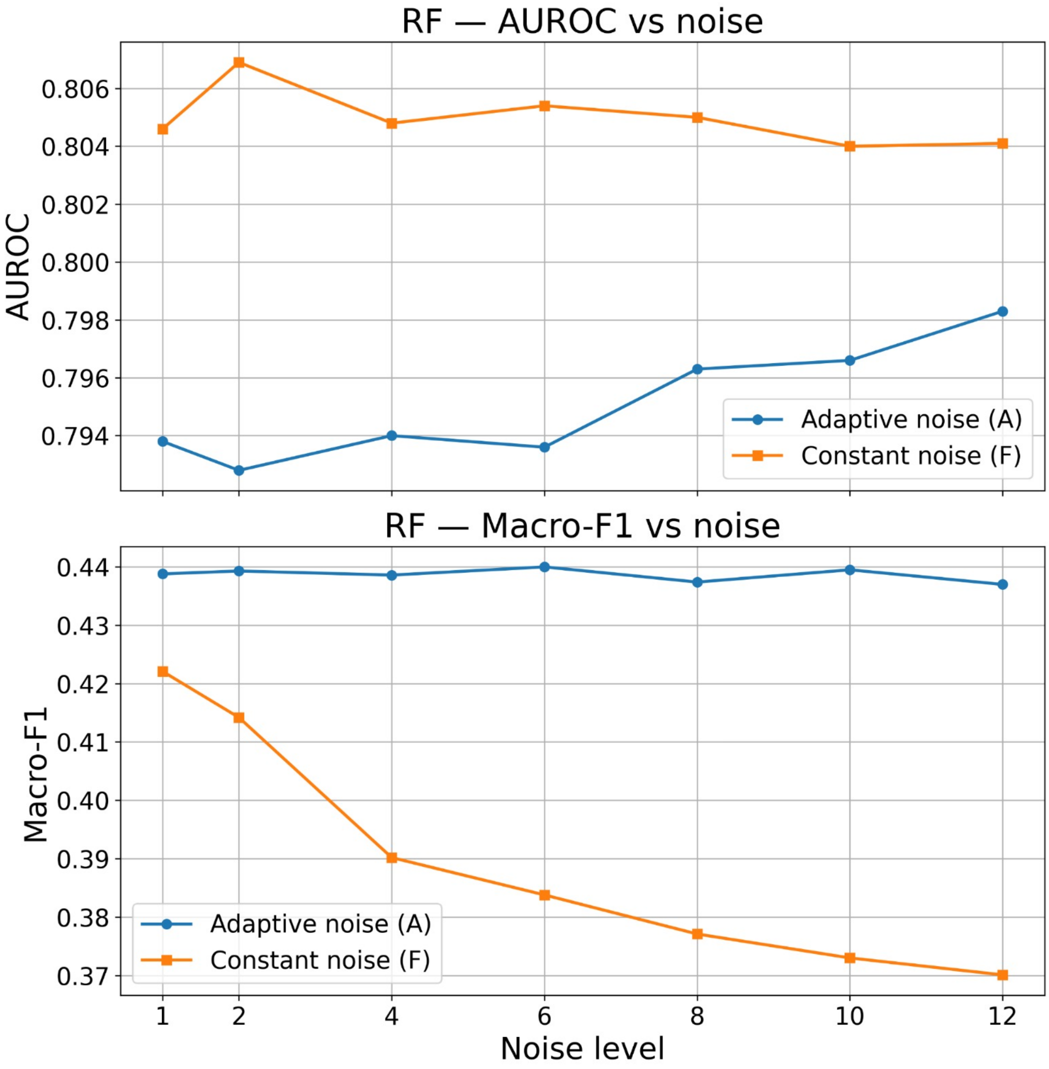


Figure 9: Random Forest: Effect of oversampling noise level on Macro-F1 and AUROC. Adaptive noise consistently improves Macro-F1 as noise level increases, while constant noise degrades performance. Optimal configuration: 10% adaptive noise.

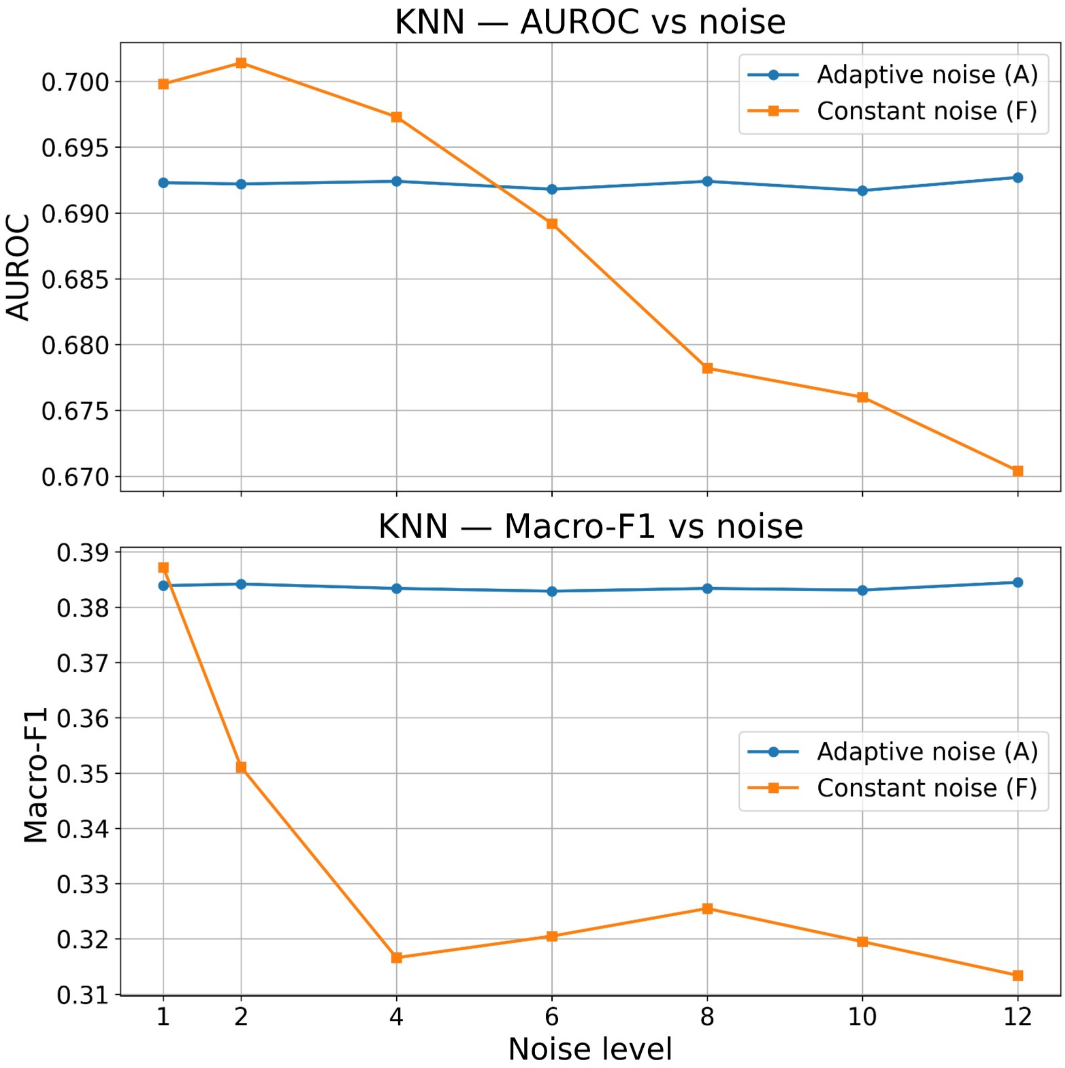


Figure 10: KNN: Effect of oversampling noise level on Macro-F1 and AUROC with distance weighting. Adaptive noise shows stable Macro-F1 performance, while constant noise severely degrades both metrics. Best performance at 12% adaptive noise.

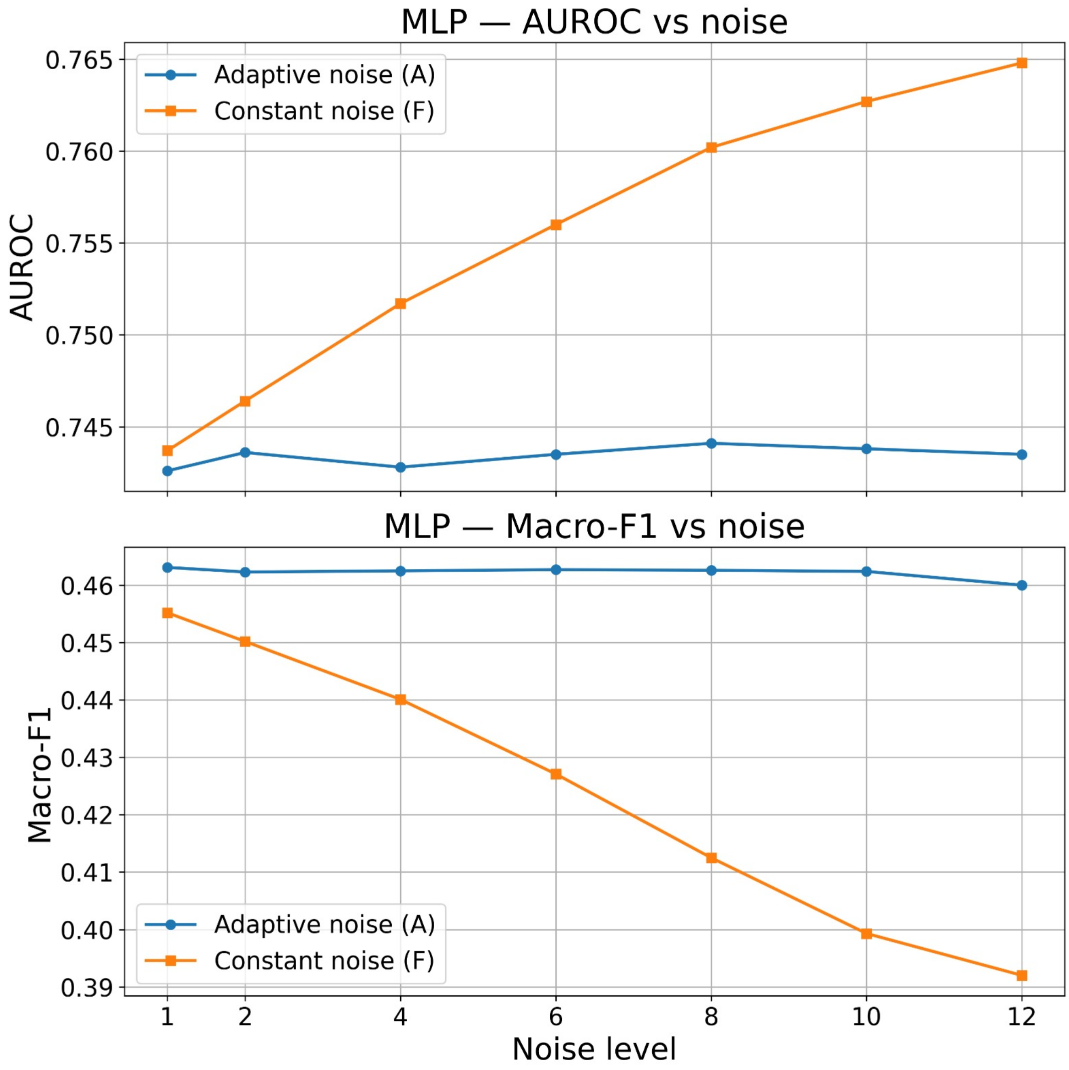


Figure 11: MLP: Effect of oversampling noise level on Macro-F1 and AUROC. Constant noise improves AUROC but degrades Macro-F1, while adaptive noise maintains stable Macro-F1. Optimal configuration: 10% adaptive noise.

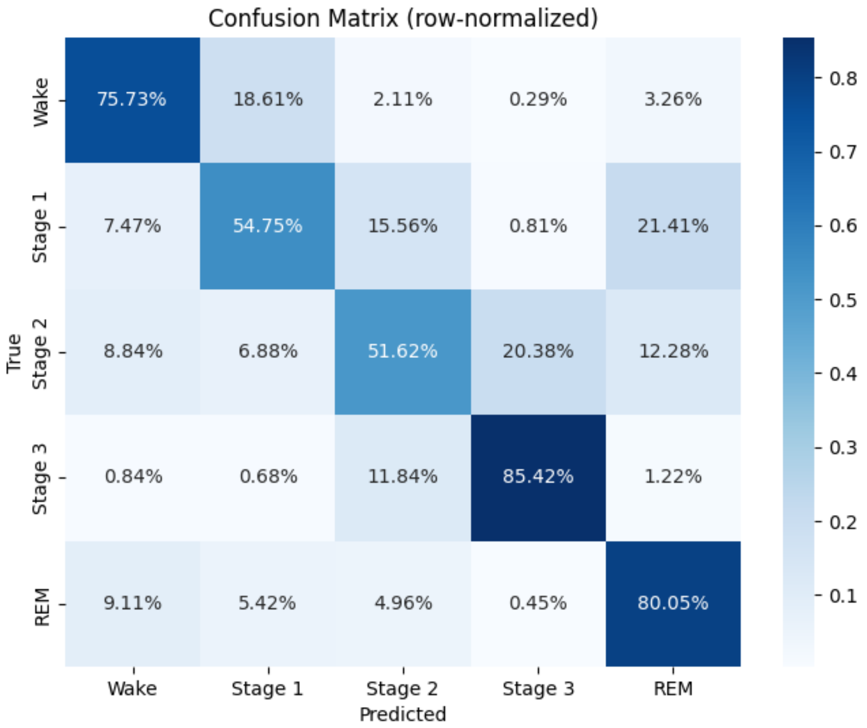


Figure 12: SleepFM sleep staging confusion matrix (0.7 accuracy).

References

[1] American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events.* Darien, IL, 2012.

[2] Bingbing Bai, Jialin Liang, Guanhua Zhang, Huan Li, Kun Bai, and Fei Wang. Why attentions may not be interpretable? In *KDD 2021*, pages 25–34, 2021.

[3] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29, 2004.

[4] Geoffrey Brookshire. Data leakage in deep learning studies of translational eeg. *Frontiers in Neuroscience*, 2024.

[5] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[6] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[7] Brett Duce, Colin Rego, Jasminka Milosavljevic, and Craig Hukins. The aasm recommended and acceptable eeg montages are comparable for the staging of sleep and scoring of eeg arousals. *Journal of Clinical Sleep Medicine*, 10(7):803–809, 2014.

[8] Sahibsingh A. Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(4):325–327, 1976.

[9] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from Imbalanced Data Sets.* Springer, 2018.

[10] Minfang Fu, Yi Wang, Zhuo Chen, Jianqing Li, Fei Xu, Xin Liu, and Fengzhen Hou. Deep learning in automatic sleep staging with a single channel electroencephalography. *Frontiers in Physiology*, 12, 2021.

[11] Ary Goldberger, Luis Amaral, Leon Glass, Jeffrey Hausdorff, Plamen Ch. Ivanov, Roger Mark, et al. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[12] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[13] F. Holm et al. Optimized inter-scorer agreement in sleep stage classification using deep learning. *Sleep Medicine*, 2024.

[14] Robert W. Homan. The 10-20 electrode system and cerebral location. *American Journal of EEG Technology*, 28(4):269–279, 1988.

[15] I. A. M. Huijben, S. Overeem, M. M. Van Gilst, and R. J. G. Van Sloun. Attention on sleep stage specific characteristics. In *EMBC 2024*, 2024.

[16] M. Kawai et al. Compensated integrated gradients for reliable explanation of electroencephalogram signal classification. *Brain Sciences*, 12(7):849, 2022.

[17] Ahsan Khandoker, Chandan Karmakar, Thomas Penzel, Martin Glos, and Marimuthu Palaniswami. Investigating relative respiratory effort signals during mixed sleep apnea using photoplethysmogram. *Annals of Biomedical Engineering*, 41:2101–2111, 2013.

[18] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV 2017*, 2017.

[20] Ravi K. Malhotra and Alon Y. Avidan. Sleep stage scoring. In *Atlas of Sleep Medicine*, pages 125–163. Springer International Publishing, third edition, 2023.

[21] K. Markov, M. Elgendi, V. Birrer, et al. Interpretable feature-based machine learning for automatic sleep detection using photoplethysmography. *npj Biosensing*, 2(24), 2025.

[22] M. Perslev et al. U-sleep: Resilient high-frequency sleep staging. *npj Digital Medicine*, 4(1), 2021.

[23] H. Phan et al. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 2022.

[24] M. Radha, G. Garcia-Molina, M. Poel, and G. Tononi. Comparison of feature and classifier algorithms for online automatic sleep staging based on a single eeg signal. In *EMBC 2014*, pages 1876–1880, 2014.

[25] M. Rosenblatt et al. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, 2024.

[26] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), 2015.

[27] Mohammadreza Tashakori, Minttu Rusanen, Tuomas Karhu, Ludger Grote, Rajendra K. Nath, Timo Leppänen, and Sami Nikkonen. Interhemispheric differences of electroencephalography signal characteristics in different sleep stages. *Sleep Medicine*, 117:201–208, 2024.

[28] Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore, Gauri Ganjoo, Emmanuel Mignot, and James Zou. Sleepfm: Multi-modal representation learning for sleep across brain activity, ecg and respiratory signals, 2024.

[29] Rahul Thapa, Magnus R. Kjær, Bryan He, Ian Covert, Hyatt Moore, Usman Hanif, et al. A multimodal sleep foundation model developed with 500k hours of sleep recordings for disease predictions, 2025. medRxiv [Preprint].

[30] F. Vaquerizo-Villar et al. An explainable deep-learning model to stage sleep states in children and propose novel eeg-related patterns in sleep apnea. *Computers in Biology and Medicine*, 165:107419, 2023.

[31] R. S. Vulcan, S. André, and M. Bruyneel. Photoplethysmography in normal and pathological sleep. *Sensors*, 21(9):2928, 2021.