Automatic Sleep Disorder Classification

Team:
Popa Stefan Andrei
Jonas Dauksa
Daniel Skuczi
Eduard Levinschi
Yannick Brackelaire
Aedem Bangerter

Supervisors:
Dr. Iris Huijben
Dr. Pietro Bonizzi

Coordinator:
Dr. Menica Dibenedetto

Maastricht University

Research Question

Can pre-trained foundation model representations (SleepFM) generalize better than handcrafted statistical features for sleep disorder diagnosis in severely imbalanced clinical cohorts, and which algorithmic strategies are most effective for handling class imbalance in this setting?

The problem

- The prevalence of sleep disorders has increased significantly[1],[2],[3]
- Access to sleep labs and specialists is limited

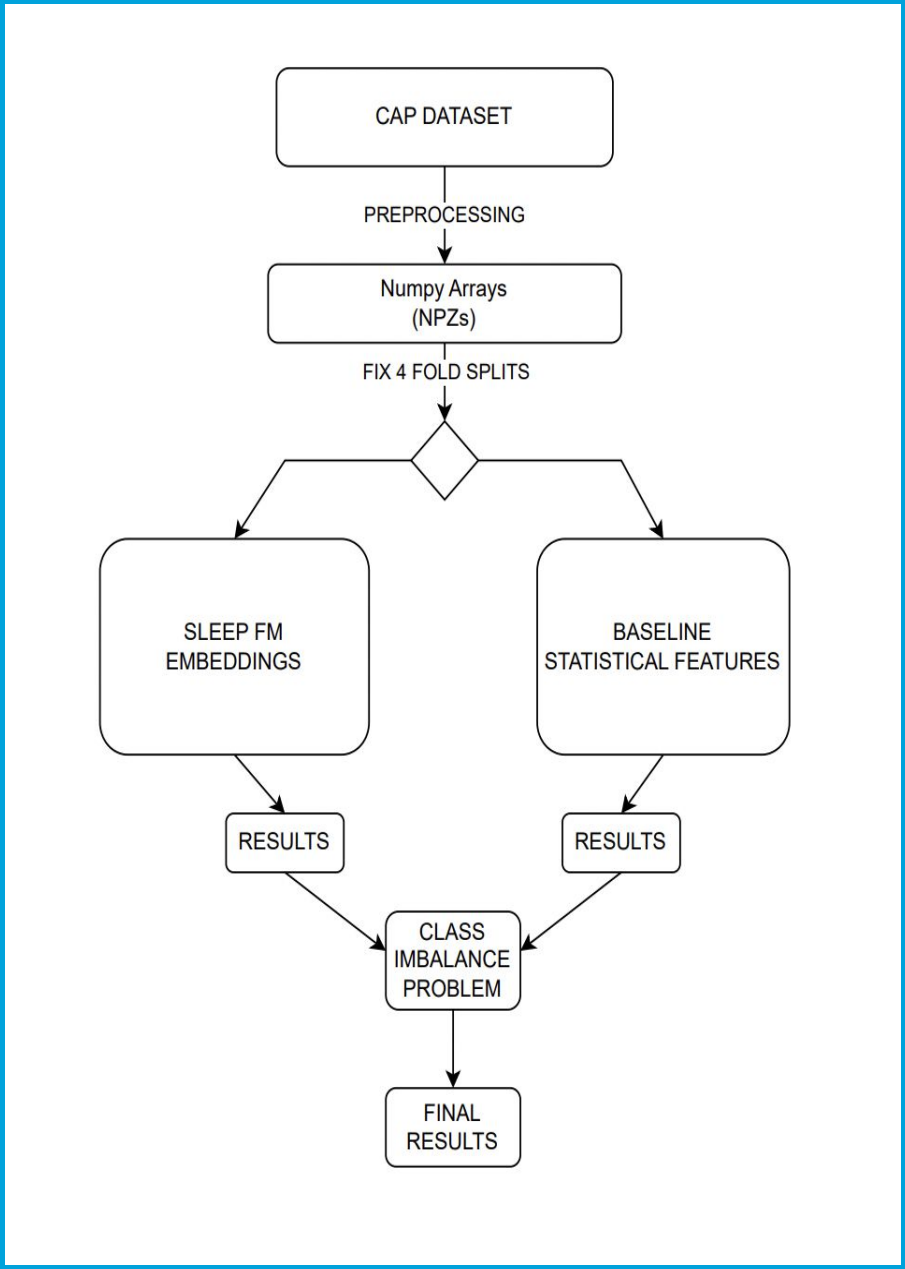
Our goal:

Automatically classify sleep disorders from EEG data.

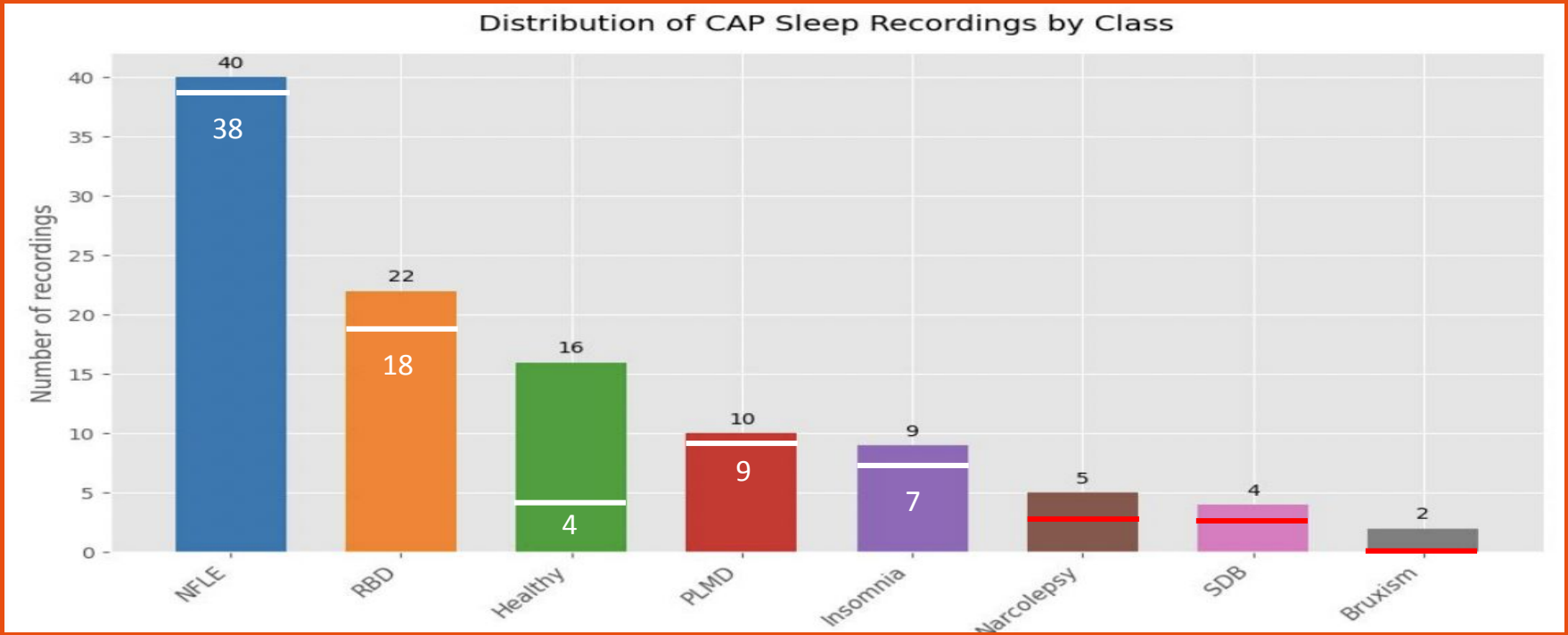
Approach:

Compare classical ML with handcrafted features against a foundation model.

Overview


CAP DATASET
PREPROCESSING
Numpy Arrays
(NPZs)
FIX 4 FOLD SPLITS
SLEEP FM
EMBEDDINGS
BASELINE
STATISTICAL FEATURES
RESULTS
RESULTS
CLASS
IMBALANCE
PROBLEM
FINAL
RESULTS

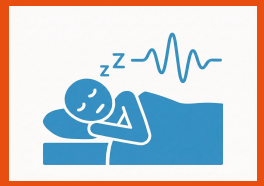
CAP Sleep Dataset[4],[5]



Due to channel availability, the following classes were kept:

NFLE
Nocturnal Frontal Lobe Epilepsy

RBD
REM Behavior Disorder



Control
Healthy

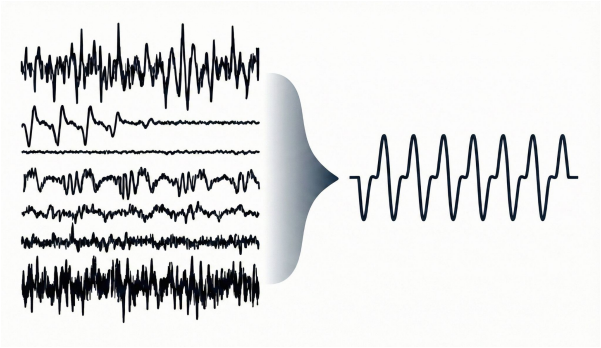
PLM
Periodic Limb Movements

Insomnia

Maastricht University

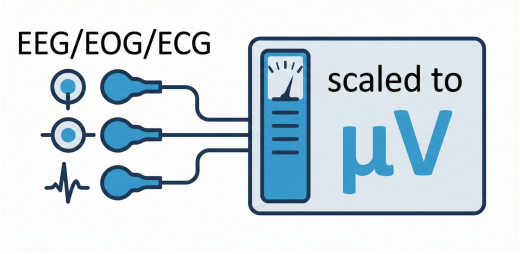
Preprocessing

Signal Standardization



- Resampling: All signals ⟶ 256 Hz
- Epoching: 30-second windows (7,680 samples)

Channel-Specific Processing



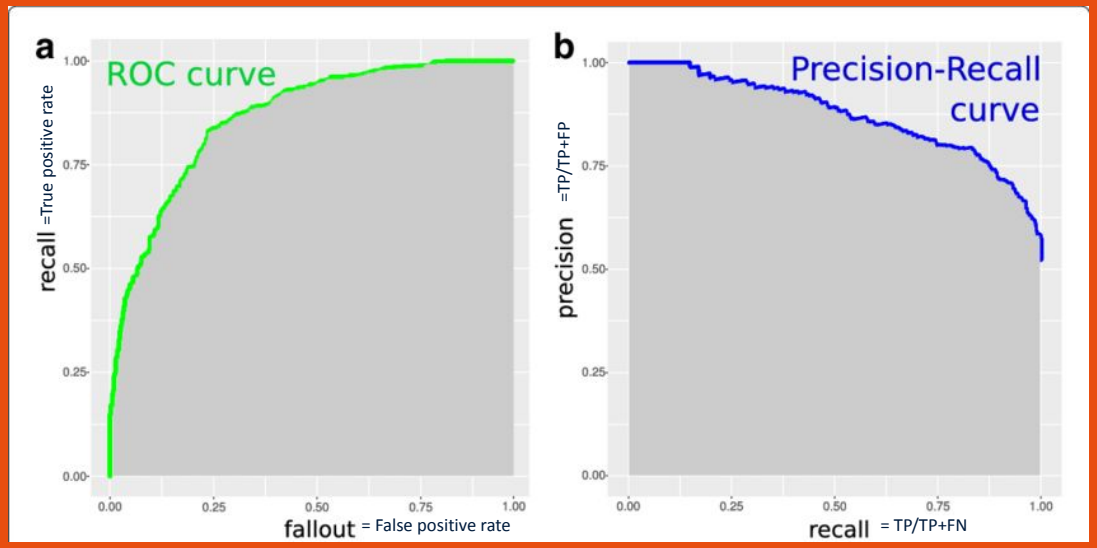
Artifact cleaning (gradient drop detection, linear interpolation)

Metrics Explanation

Macro F1-Score:
Averages Precision & Recall per class (equal weight).

Low F1: Model ignores rare classes.

High F1: Balanced detection across sleep disorders.



AUROC:
Class separation/ranking ability.
High AUROC: Good classification

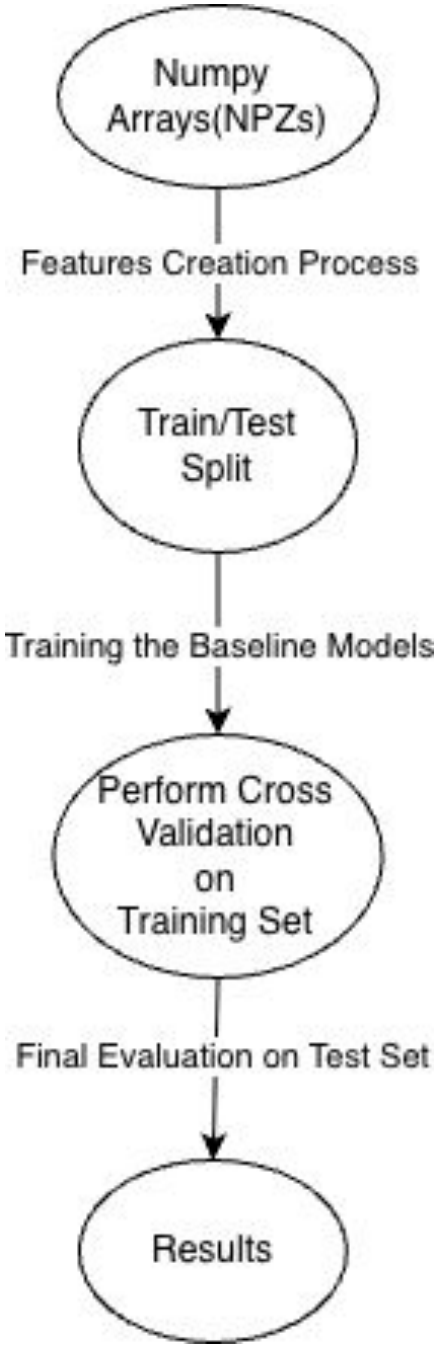
Can mask poor performance on minority classes.

AUPRC:
Precision-Recall trade-off (emphasizing minority).
Low AUPRC: Many false positives for rare disorders

Baseline Approach Structure

Objectives

- Build classical ML baselines for sleep disorder classification
- Provide a strong comparison point for deep models



Features Creation

Time-Domain Features: capture amplitude and variability of the signal

- Root Mean Squared
- Standard Deviation
- Length of the Night

for each channel
except
photoplethysmography &
oxygen saturation

Frequency-Domain Features: sleep disorders strongly affects frequency content

- Computing PSD(power spectral density) using Welch's method
- Delta(0.5-4), Theta(4-8), Alpha(8-12), Sigma(12-15), Beta(15-30) Hz
- For each band, the power inside is computed and normalized by total power

Statistical Features Dataset

- Combine all time-domain and frequency-domain features in order to create one vector per channel
- Concatenate feature vectors from all channels
- Repeat the process for every EEG segment

Train/Test Split

- Split the dataset into train and test set
- Split was performed after feature creation

- 4 splits cross validation was performed in the end for the final experiments

Stratified Splitting

Why?

- Dataset is class-imbalanced
- Random splits could remove rare classes from the test set
- Stratification ensures fair evaluation

How?

- Used stratified splitting
- Ensured that minority classes are represented
- It was done by class label

Random Forest Model

- Configuration: number of trees(400) which is great number for stability, random seed was fixed for reproducibility results

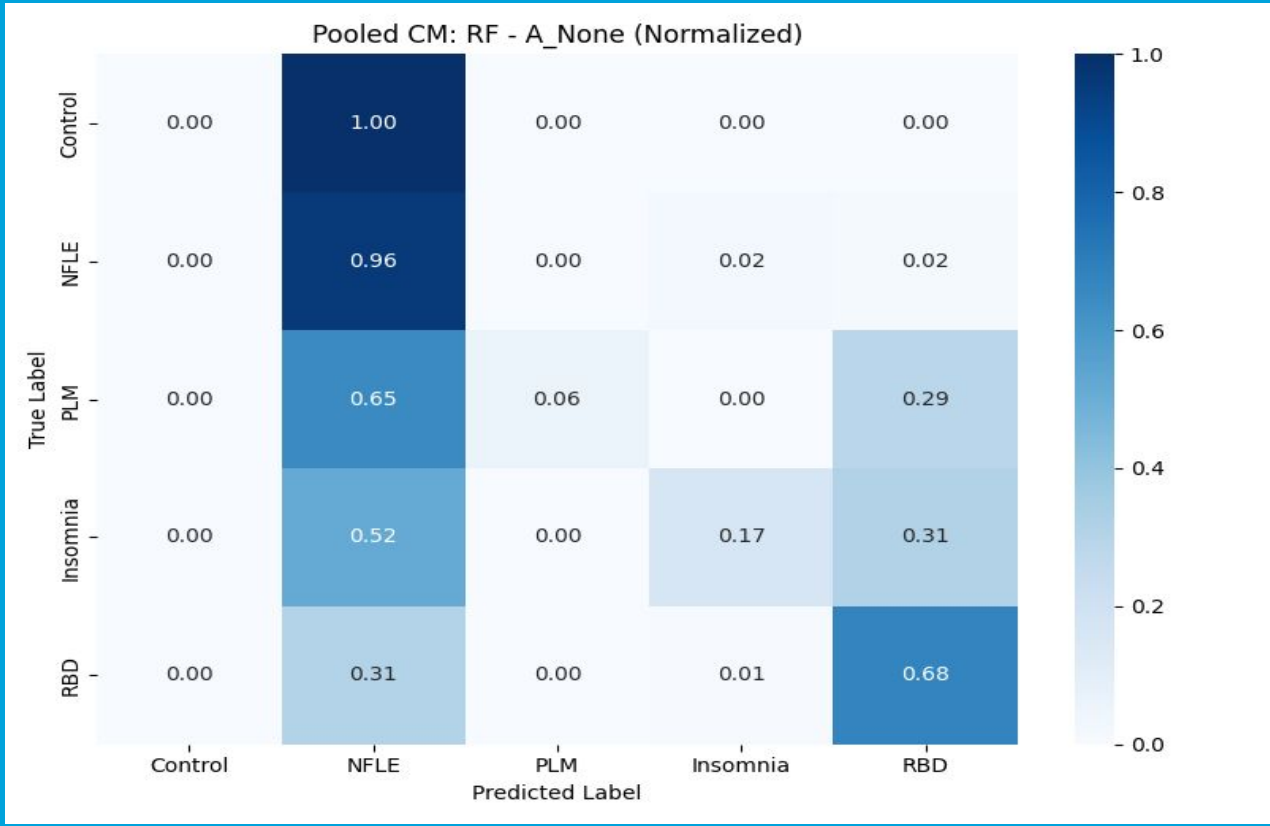
- Random Forest do not rely on distances, so feature scaling was not a requirement

KNN Model

- Configuration: Selecting the K nearest-neighbor (which was chosen from 3 to 21)
- Small K -> sensitive to noise, Large K -> smoother
- Feature Scaling: we applied standardization(zero mean, unit variance)
- Distance calculation require comparable feature scales

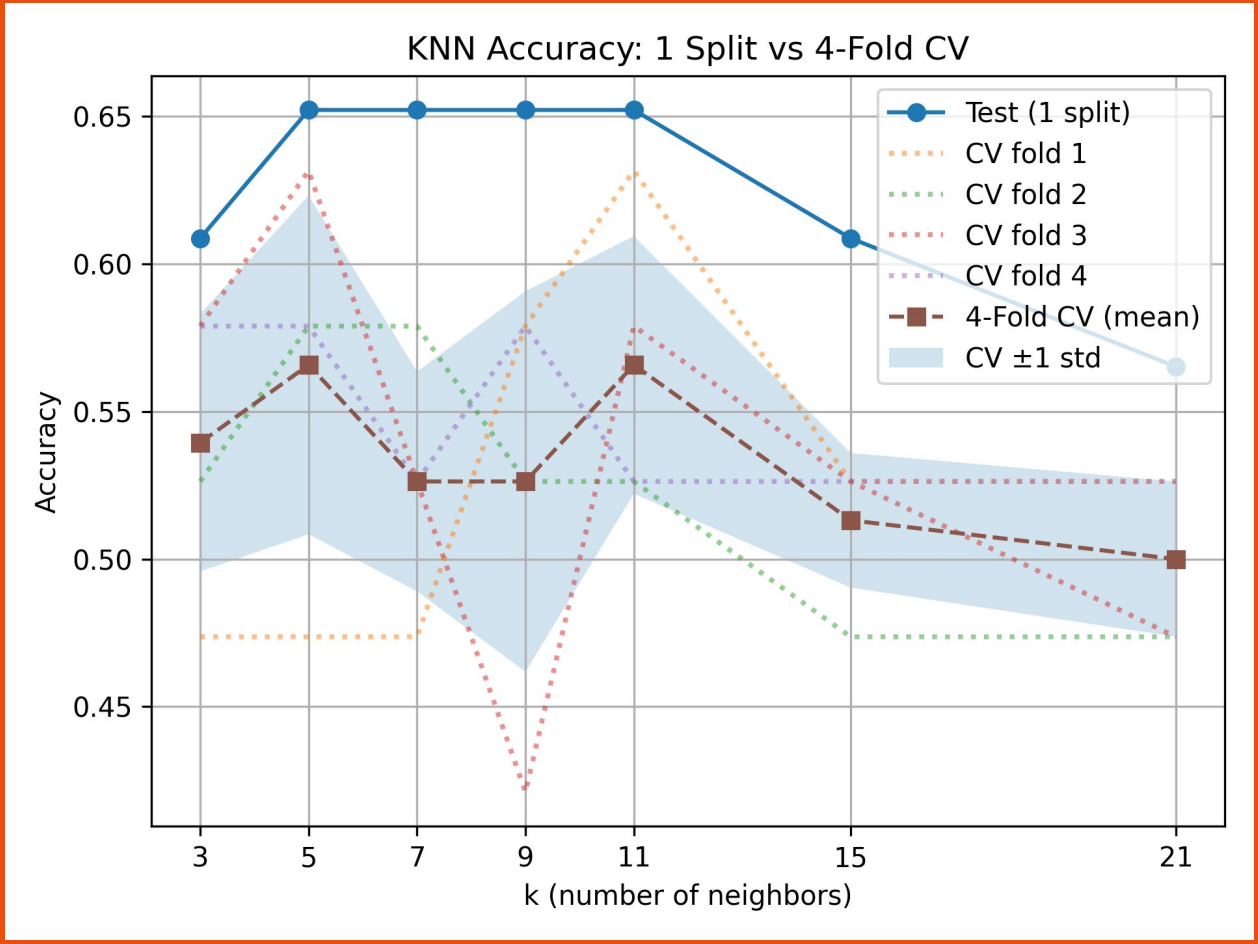
Random Forest + CV Results

Accuracy	0.7764
Macro F1-Score	0.3456



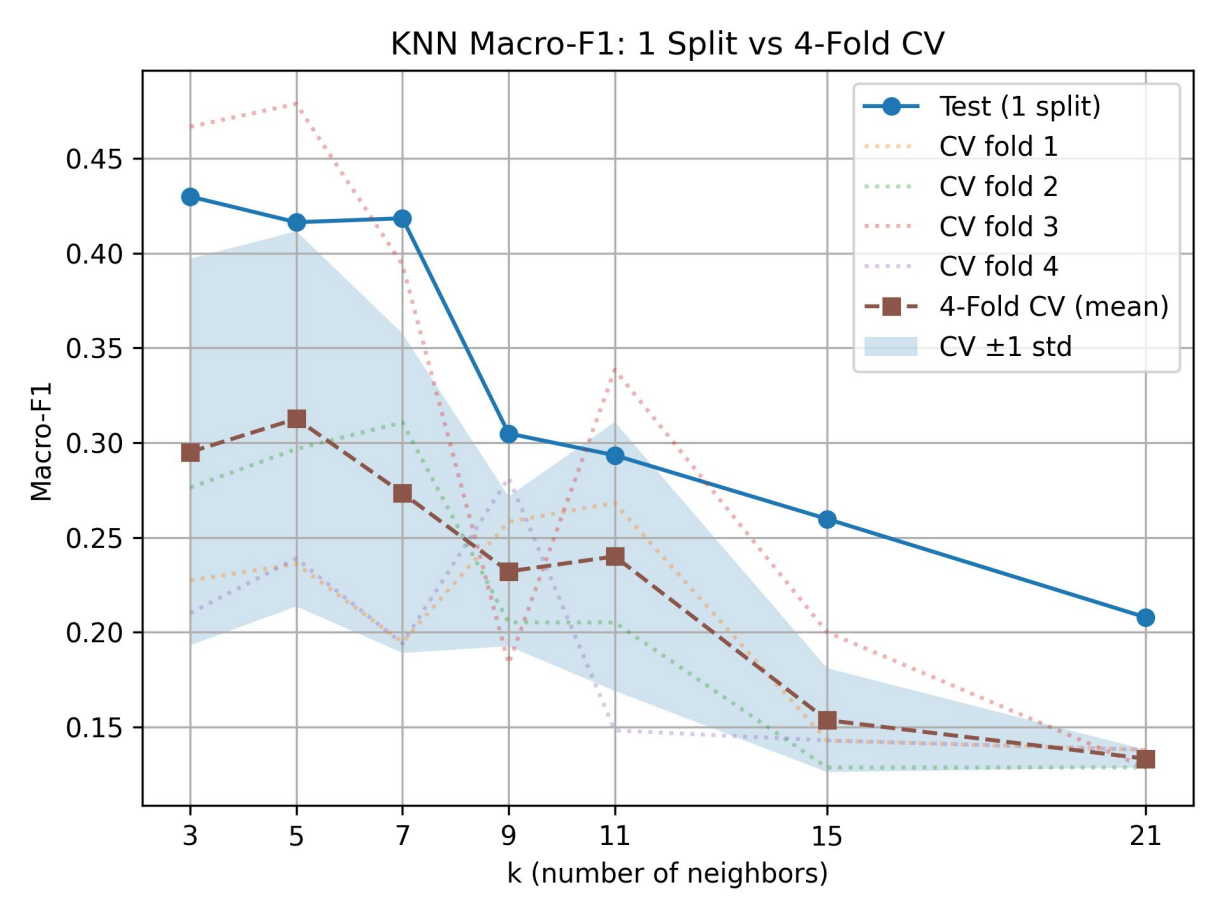
KNN(without CV) vs KNN(CV) Results

- Cross-validation gives a lower but more reliable estimate because it evaluates the model on multiple different splits instead of a single potentially lucky one



KNN(without CV) vs KNN(CV) Results

- Larger k values further bias predictions toward majority classes, reducing balanced performance
- Macro-F1 decreases under cross-validation



What is SleepFM?

Pre-trained **foundation model**[6] for sleep stage classification.

Trained on Stanford Sleep Clinic dataset (over 100,000 hours of polysomnography data).

Self-supervised contrastive learning with 19 channels over 3 modalities.

Can these pre-trained representations generalize to sleep disorders on our small CAP dataset?

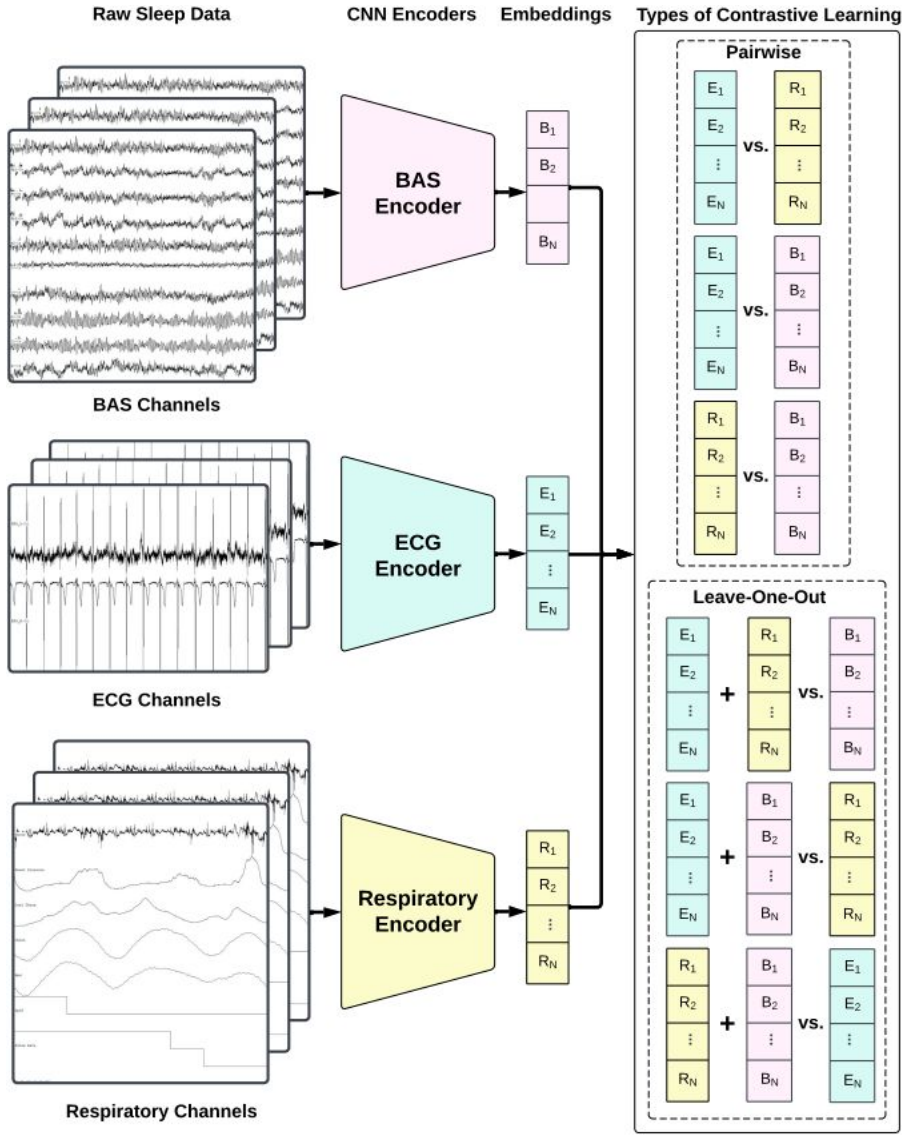
SleepFM Architecture

Three EfficientNet-based[7],[8] encoders:

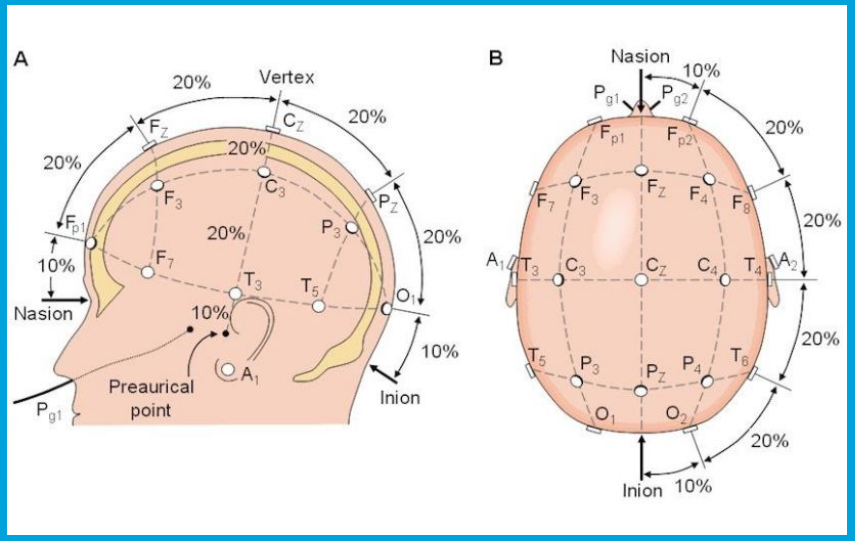
EEG/EOG (BAS)
Input: 5 channels
Captures sleep staging patterns: sleep spindles, K-complexes, REM eye movement

Cardiac
Input: 1 channel
Captures heart rate variability

Respiratory
Input: 3 channels
Captures breathing patterns, apnea events



Channel mapping



SleepFM Expectation

Respiratory:

CHEST, SaO_2, ABDOMEN

Brain Activity Signals & EOG:

C3-A2, C4-A1, O1-A2, O2-A1, E1-A2

Cardiac: ECG1-ECG2

Substituted Mechanical with Optical

Spatial locations, physiologically similar

No changes needed

CAP Available Data

Respiratory:

Photoplethysmography, SaO_2

Brain Activity Signals & EOG:

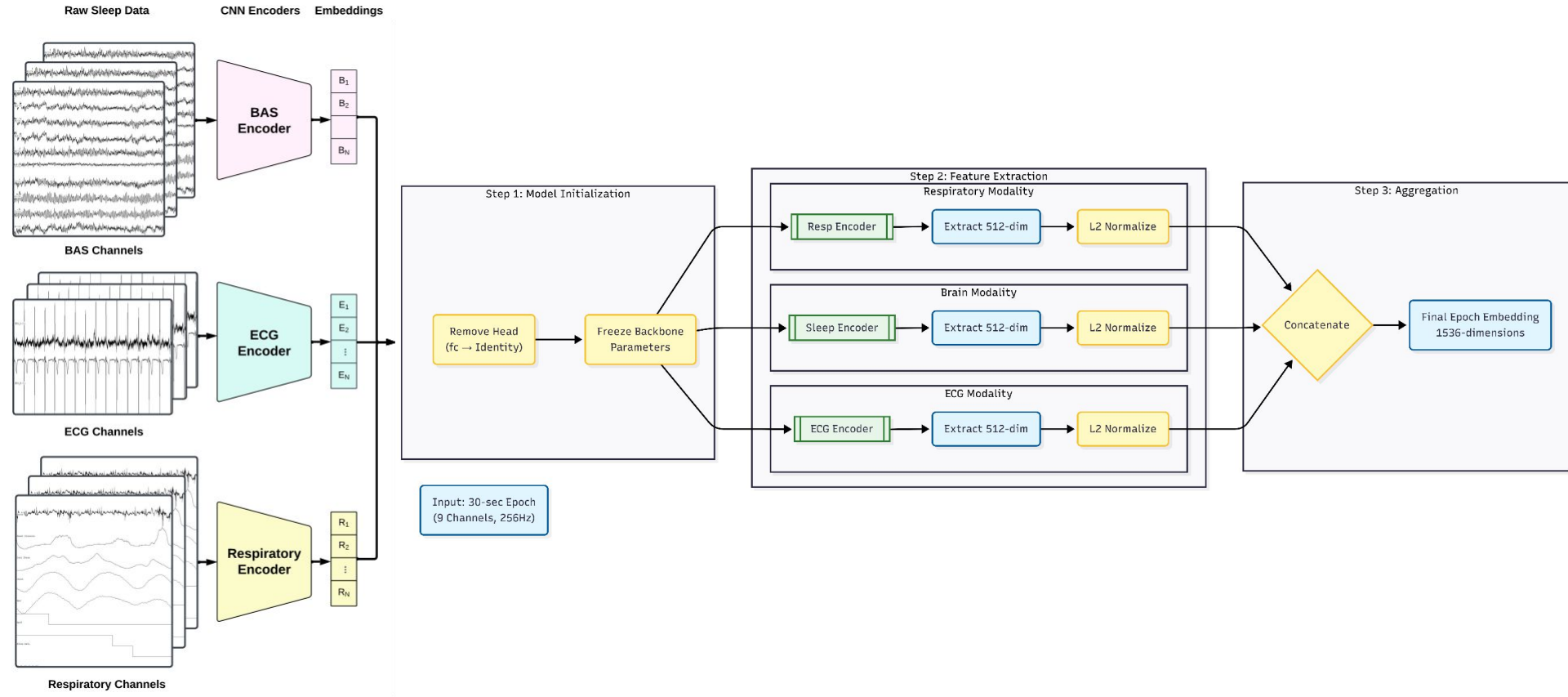
F4-C4, C4-A1, C4-P4, P4-O2, ROC-LOC

Cardiac: ECG1-ECG2

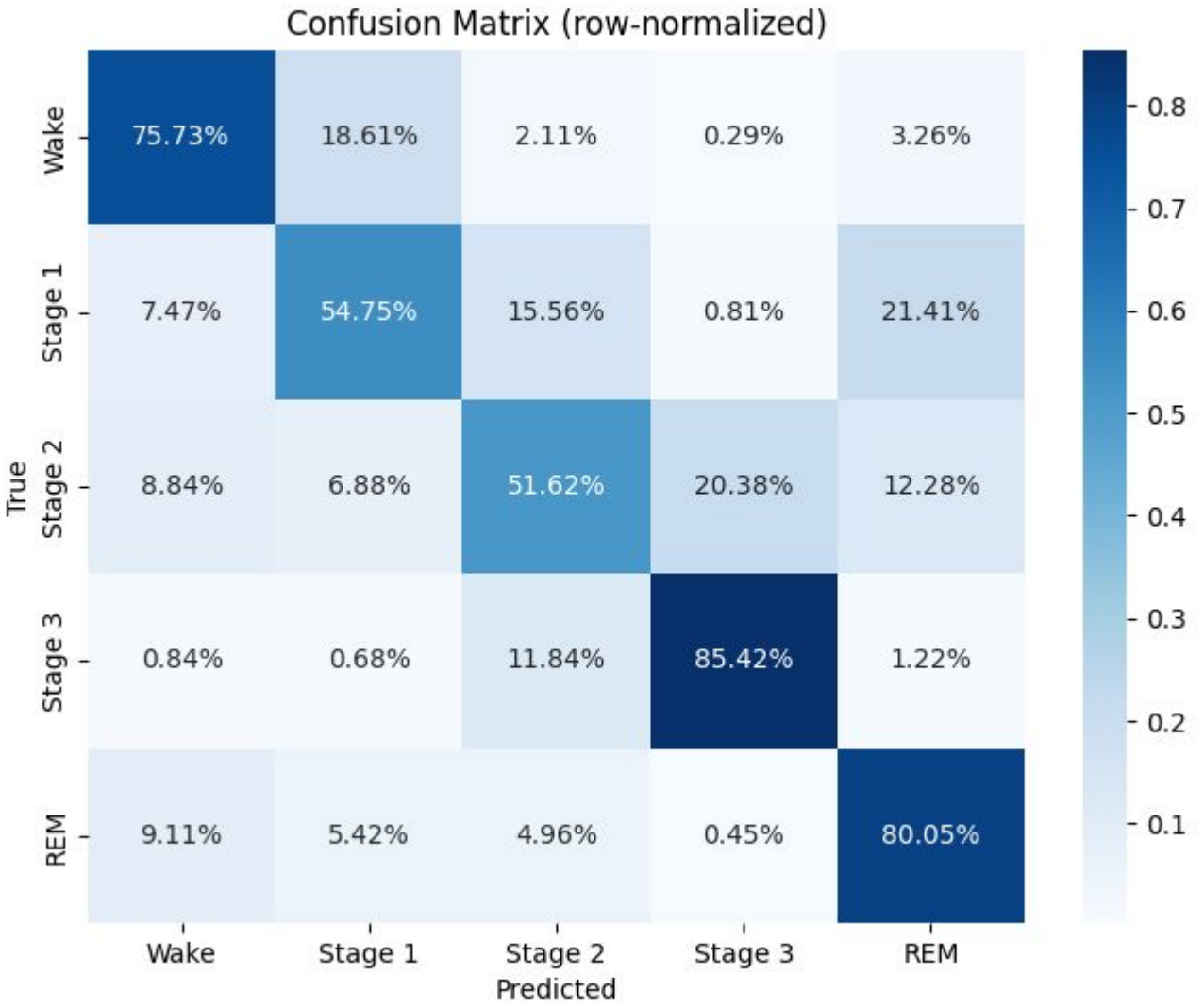
Maastricht University

The 10–20 International system seen from left (**a**) and above the head (**b**). The letters F, T, C, P, and O stand for frontal, temporal, central, parietal, and occipital lobes. (Malmivuo and Plonsey (1995), WEB version) Campisi, P., Maiorana, E. (2025). EEG Biometrics. In: Jajodia, S., Samarati, P., Yung, M. (eds) Encyclopedia of Cryptography, Security and Privacy. Springer, Cham. https://doi.org/10.1007/978-3-030-71522-9_1485

Embedding Generation



SleepFM Sleep Staging (CAP)

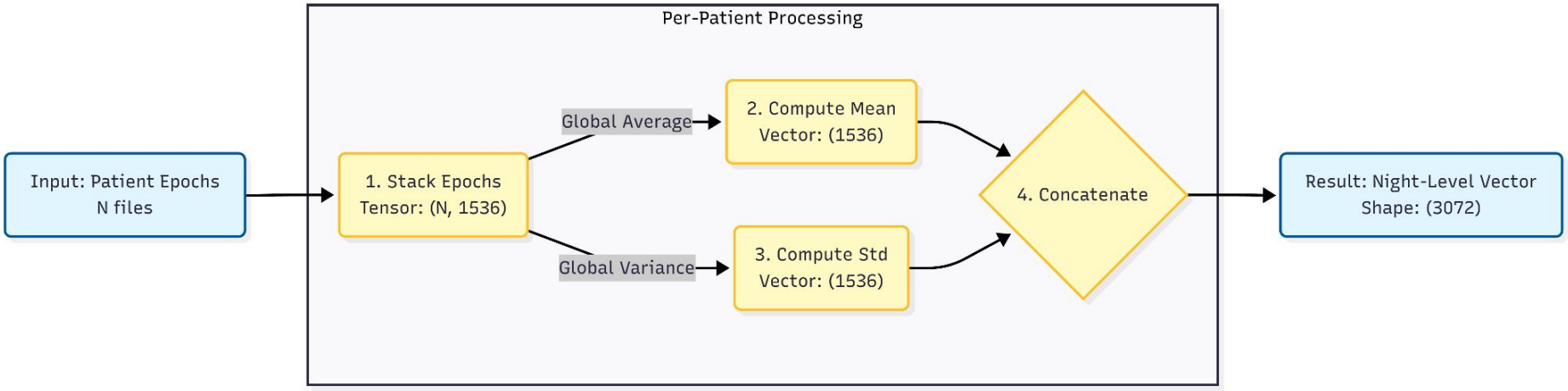


Classification Report	
Accuracy	70 %
Macro F1	64 %
AUROC	91 %
AUPRC	69 %

Epoch-Level ⟶ Night-Level Aggregation

Problem: One patient has ~1000 of 30-second epochs

Solution: Aggregate to single night-level feature vector



SleepFM Sleep Disorder Classification

4-Fold Stratified Cross-Validation:

- 50 seeds.
- No data leakage.
- One-vs-Rest Macro Averaging
- 95% confidence intervals.
- Consistency: every fold pre-defined

Models:

1. Logistic Regression (L1 with C=0.1, max iterations 2000)
2. Random Forest (400 trees, no max depth)
3. KNN (z-score normalization, k=5, distance weighting)

Imbalance handling scenarios:

A. No intervention (baseline)

B. Focal Loss (y=2.0)

C. Class Weights

D. Oversampling with Adaptive Noise (1-12%, scaled by σ of each feature)

Class Imbalance Solution

Theory and Methods [9,10]

- Oversample rare disorders:
  - -> Without additive noise
  - -> With some additive noise

  or, use a class-aware subject sampler

- Loss-based class weighting (RF)
- Distance-weighting (KNN)

- Rare-class removal

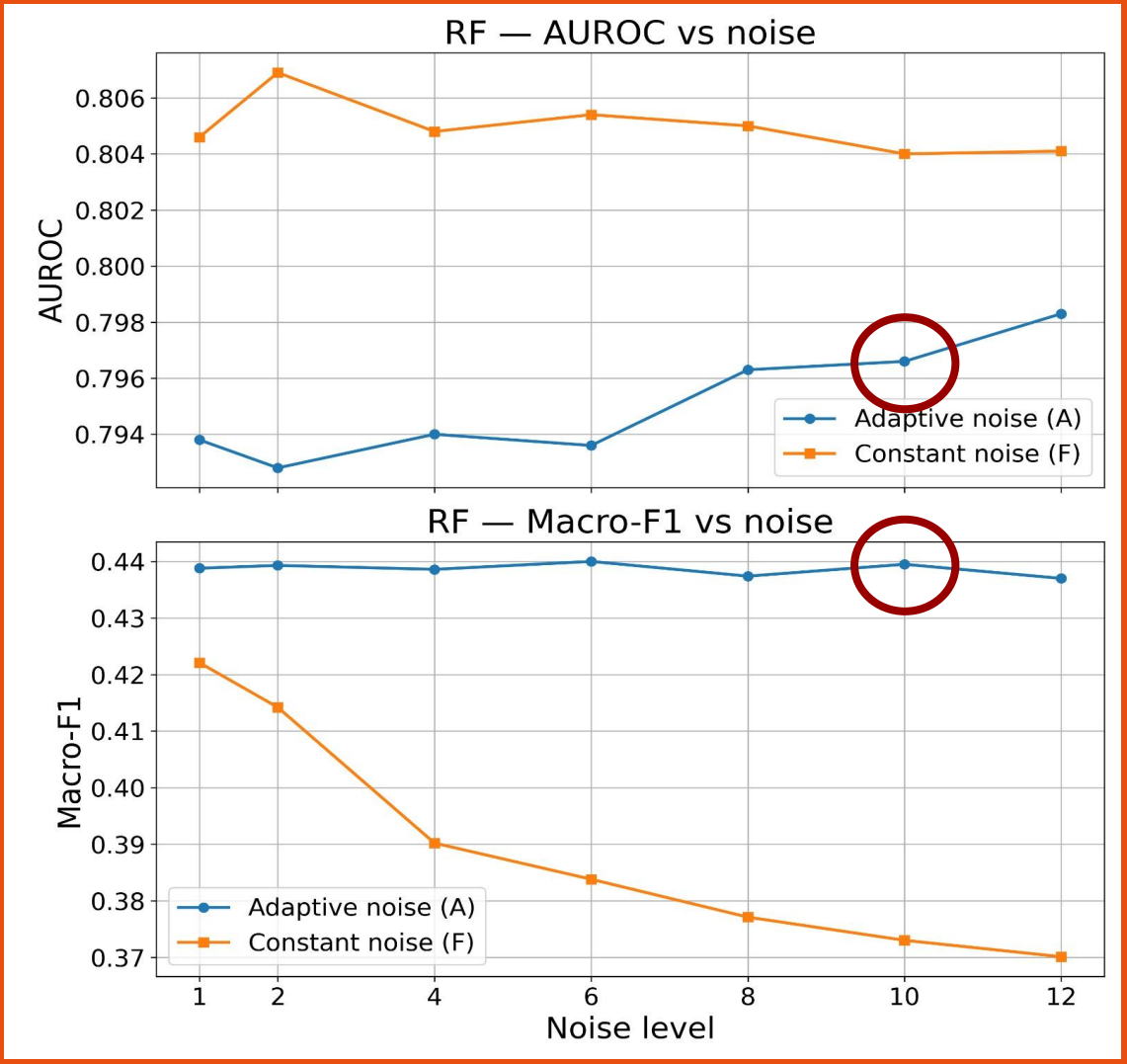
1. Alberto Fernandez, Salvador Garcia, Mikel Galar, Ronaldo C. Prati and Francisco Herrera. Learning from Imbalanced Data Sets. Springer, 2018
2. Guillaume Lemaitre, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17):1–5, 2017

Experiments on our baseline

Comparison Table of RF's Models

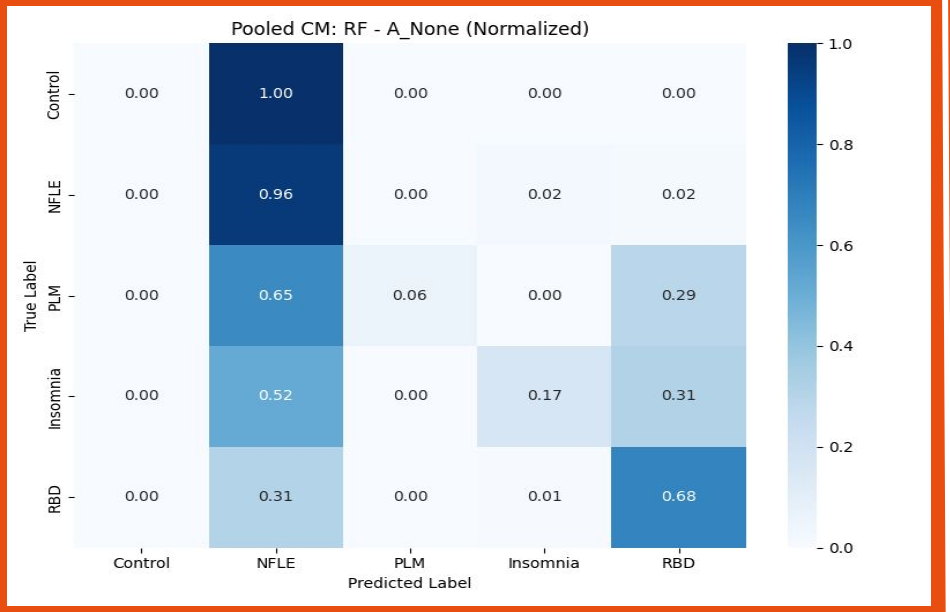
Scenario	F1 Score (95% CI)	AUROC (95% CI)	AUPRC (95% CI)
Random Forest Basic	0.345 (0.335-0.356)	0.786 (0.779-0.794)	0.613 (0.603-0.622)
Random Forest + Lose Weighting	0.356 (0.342-0.370)	0.797 (0.791-0.803)	0.638 (0.628-0.649)
Random Forest + Oversampling	0.438 (0.425-0.450)	0.796 (0.789-0.805)	0.704 (0.694-0.715)
Random Forest + Oversampling (10% noise)	**0.439 (0.426-0.453)**	**0.798 (0.792-0.805)**	**0.705 (0.697-0.713)**

Graphs of Performance over Noise

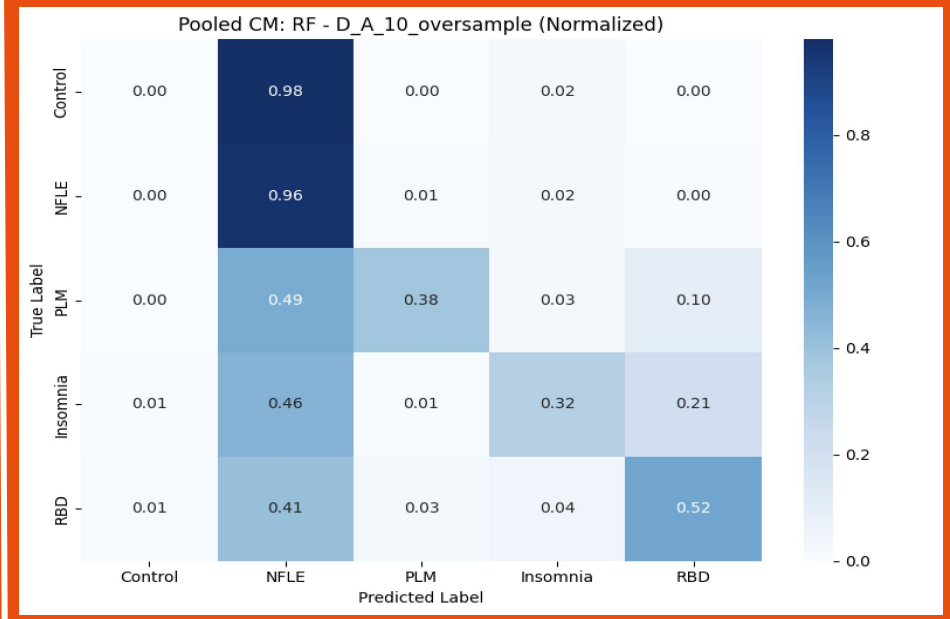


Confusion Matrices

Random Forest Basic



Random Forest Oversampling(10%)



Maastricht University

Comparison Table of KNN's

Scenario	F1 Score (95% CI)	AUROC (95% CI)	AUPRC (95% CI)
KNN Basic	0.245 (0.235-0.256)	0.667 (0.656-0.675)	0.388 (0.379-0.397)
Distance -Weighted KNN	0.320 (0.310-0.329)	0.696 (0.688-0.703)	0.483 (0.472-0.494)
KNN + oversampling	0.384 (0.374-0.395)	0.692 (0.686-0.670)	0.540 (0.531-0.549)
KNN + Oversampling (12% noise)	**0.385 (0.374-0.396)**	**0.693 (0.687-0.672)**	**0.540 (0.530-0.549)**

Experiments on SleepFM

Scenario	F1 Score (95% CI)	AUROC (95% CI)	AUPRC (95% CI)
Logistic Regression (baseline)	0.281 (0.273-0.290)	0.748 (0.739-0.756)	0.531 (0.520-0.543)
Logistic Regression + oversampling	0.340 (0.332-0.350)	0.754 (0.746-0.764)	0.511 (0.500-0.523)
RF (baseline)	**0.411 (0.401-0.419)**	**0.868 (0.861-0.875)**	**0.675 (0.663-0.687)**
RF + Class-weighted loss	0.343 **(0.334-0.351)**	0.822 (0.815-0.829)	0.698 (0.686-0.710)
RF + oversampling	0.402 (0.394-0.408)	0.861 (0.854-0.868)	0.678 (0.660-0.689)
KNN (baseline)	0.360 (0.353-0.366)	0.759 (0.750-0.768)	0.478 (0.465-0.489)
KNN + oversampling	**0.465 (0.460-0.471)**	**0.801 (0.793-0.809)**	0.756 (0.746-0.767)

Result and Discussion

Approach	Model	F1 Score (95% CI)	AUROC (95% CI)
SleepFM Embeddings	KNN (oversampling)	0.465 (0.460-0.471)	0.801 (0.793-0.809)
Handcrafted Features	RF (oversampling)	0.439 (0.426-0.453)	0.798 (0.792-0.805)
SleepFM Embeddings	RF (none)	0.411 (0.401-0.419)	0.868 (0.861-0.875)
Handcrafted Features	KNN (oversampling)	0.385 (0.374-0.396)	0.693 (0.687-0.672)

RF (**none**) ➡ high performance ➡ SleepFM embeddings already separate the disorders well

Oversampling is always the most effective strategy to handle class imbalance

! Results correspond to particular data set, models and parameters !

Explainability (Future Work)

- From “black box” to clinical insights

- 3 main steps:
  -> localization
  -> validation
  -> discovery

Validation

- Saliency Mapping
- Goal: Ensure AI align with experts

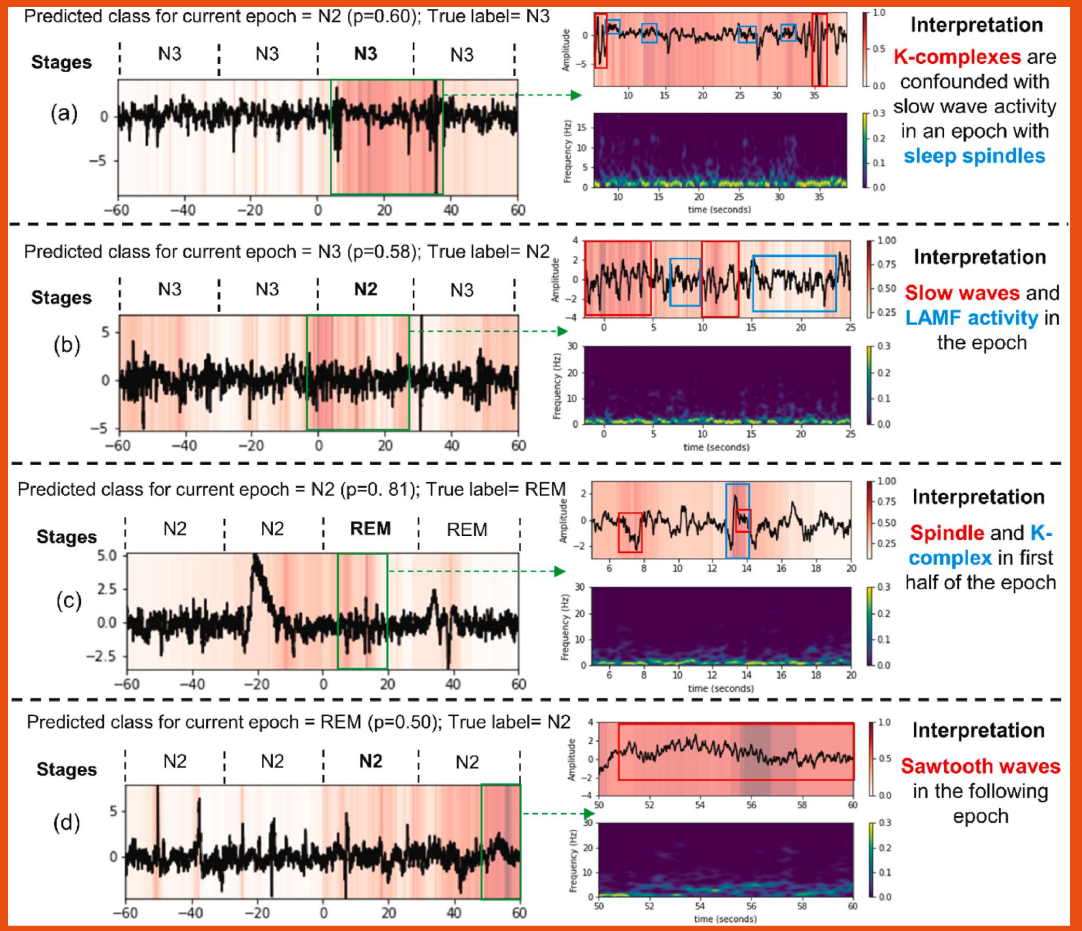


Image Source: Vaquerizo-Villar et al. (2023), *Computers in Biology and Medicine*.

Discovery

- Uncover hidden biomarkers
- Predictive patterns before symptoms
- Goal: Personalised monitoring

Conclusion

- Key Result: SleepFM + Random Forest achieved AUROC = 0.87.

- Core Advantages:
  - Superior latent space separability.
  - Reduced need for heavy rebalancing.
  - Effective generalization via careful channel substitution.

References

[1] Boers, E., Barrett, M. A., Benjafield, A. V., Barnet, J. H., Ravelo, L. A., Kaye, L., Cistulli, P. A., Pépin, J. L., Armitstead, J., Sterling, K. L., Nunez, C. M., Peppard, P. E., & Malhotra, A. (2025). Projecting the 30-year burden of obstructive sleep apnoea in the USA: a prospective modelling study. *The Lancet. Respiratory medicine*, *13*(12), 1078–1086. https://doi.org/10.1016/S2213-2600(25)00243-7

[2] Garland, S. N., Rowe, H., Repa, L. M., Fowler, K., Zhou, E. S., & Grandner, M. A. (2018). A decade's difference: 10-year change in insomnia symptom prevalence in Canada depends on sociodemographics and health status. Sleep health, 4(2), 160–165. https://doi.org/10.1016/j.sleh.2018.01.003

[3] Adjaye-Gbewonyo D, Ng AE, Black LI. Sleep difficulties in adults: United States, 2020. NCHS Data Brief, no 436. Hyattsville, MD: National Center for Health Statistics. 2022. DOI: https://dx.doi.org/10.15620/cdc:117490.

[4] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220. RRID:SCR_007345.

[5] MG Terzano, L Parrino, A Sherieri, R Chervin, S Chokroverty, C Guilleminault, M Hirshkowitz, M Mahowald, H Moldofsky, A Rosa, R Thomas, A Walters. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. Sleep Med 2001 Nov; 2(6):537-553.

[6] Thapa, R., He, B., Kjaer, M. R., Moore, H., Ganjoo, G., Mignot, E., & Zou, J. (2024). SleepFM: Multi-modal representation learning for sleep across brain activity, ECG and respiratory signals. arXiv preprint arXiv:2405.17766.

[7] Ouyang, D., Theurer, J., Stein, N. R., Hughes, J. W., Elias, P., He, B., ... & Albert, C. M. (2024). Electrocardiographic deep learning for predicting post-procedural mortality: a model development and validation study. The Lancet Digital Health, 6(1), e70-e78.

[8] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

[9] Guillaume Lemaitre, 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

[10] Alberto Fernandez, Salvador Garcia, Mikel Galar, Ronaldo C. Prati and Francisco Herrera. Learning from Imbalanced Data Sets. Springer

[11] Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, 57(6)]