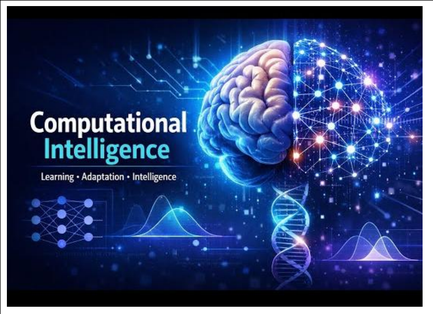


Attention-Gated 1D U-Net for Subject-Independent Sleep Spindle Segmentation
Impact of Annotation Variability and External Generalization

Waqas Ahmed[1], Keijo Haataja[2], Pekka Toivanen[3]

UNIVERSITY OF EASTERN FINLAND

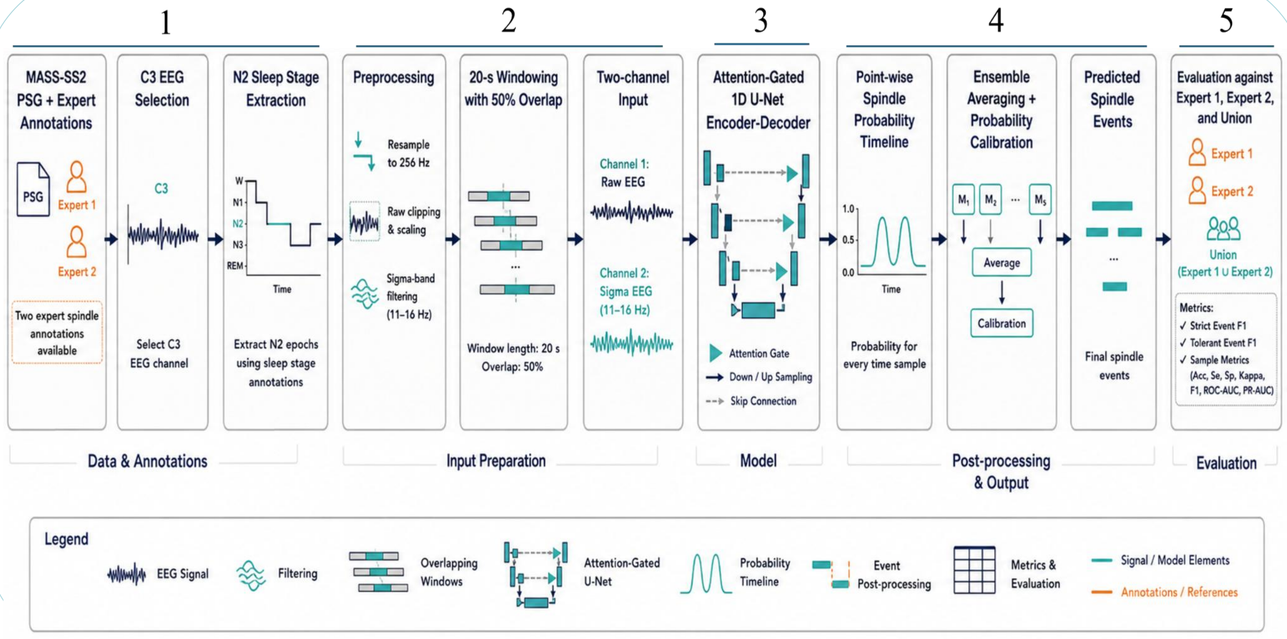
[1] Ph.D. Project Researcher, School of Computing, University of Eastern Finland, Kuopio, Finland [2] Project Manager, School of Computing [3] Professor, School of Computing

Biomedical Signal Processing | Sleep EEG | Deep Learning | External Validation

WHY THIS MATTERS

- Sleep spindles (11-16 Hz, N2 sleep) are biomarkers of sleep quality, memory consolidation, and neurological health.
- Manual spindle scoring is labor-intensive and varies substantially across experts.
- Cross-dataset deployment is difficult due to montage, sampling, cohort, and annotation differences.
- Goal: build a subject-independent, annotation-aware detector and test external zero-shot generalization.

FROM RAW EEG TO SLEEP SPINDLE SEGMENTATION



KEY RESULTS

Strict Event F1	Tolerant Event F1
0.805 ± 0.031	**0.846** ± 0.017

Sample ROC-AUC	Accuracy
0.933 ± 0.015	**0.892** ± 0.023

Subject-wise 5-fold cross-validation on MASS-SS2 using union annotations.

Main result: balanced precision and recall with robust event-level localization.

Strict P/R = 0.812 / 0.802

Tolerant P/R = 0.853 / 0.842

DATASET, MODEL SETTINGS, AND VALIDATION DESIGN

MASS-SS2 Cohort

Retained	Excluded	Original
15 subjects	**4** subjects	**19** subjects

Dual-expert analysis retained subjects with PSG, sleep stage labels, and both expert spindle annotations.

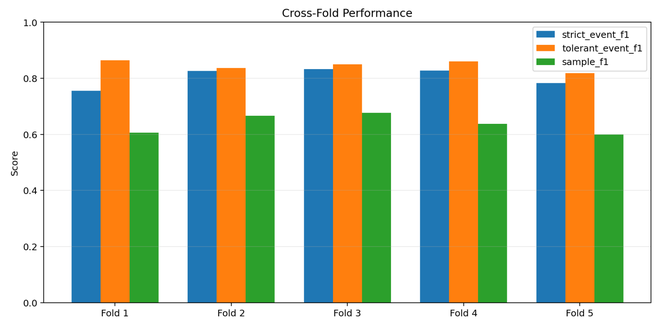
Annotation Summary

Expert 1	Expert 2	Union
10,234 events	**22,969** events	**22,576** events

Experiment Settings

- Input channel: C3 EEG; N2 sleep only.
- Resampled to 256 Hz; 20-s windows with 50% overlap.
- Dual input: normalized raw EEG + sigma-band EEG (12-16 Hz).
- Attention-gated 1D U-Net, depth 4, base channels 32.
- BCE-FPFN loss, positive-window sampling, five-seed ensemble.
- Validation-only thresholding, calibration, and event decoding.

Cross-Fold Performance



Strict F1 ranged from 0.755 to 0.833; tolerant F1 remained more stable (0.818-0.864), indicating boundary-sensitive errors.

External Zero-Shot Validation

DREAMS dataset

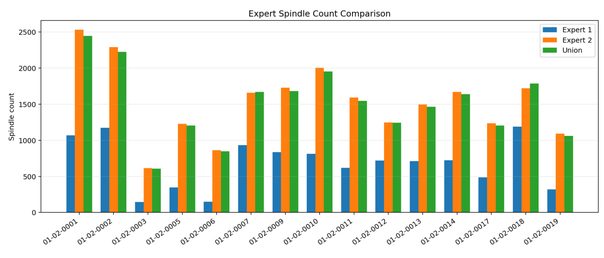
Strict Event F1	Sample ROC-AUC
0.476 zero-shot	**0.873** zero-shot

No retraining, no fine-tuning, no DREAMS-specific threshold adaptation.

- Large domain shift: montage + sampling + pathology + annotation criteria.
- Performance drop highlights the need for external validation.

ANNOTATION VARIABILITY AND ERROR ANALYSIS

Expert Annotation Variability

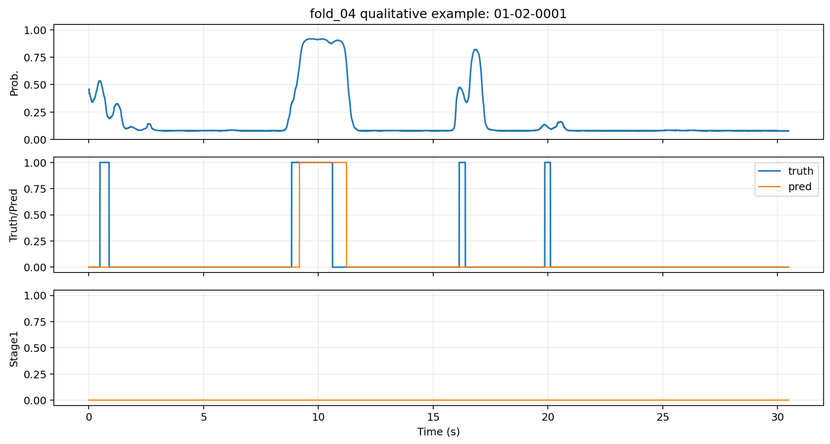


Expert 2 annotated substantially more spindles than Expert 1 across most subjects.

Against E1	Against E2	Against Union
0.563 Strict F1	**0.800** Strict F1	**0.805** Strict F1

Same predictions, different references: the chosen expert standard can change the reported F1 as much as the model design.

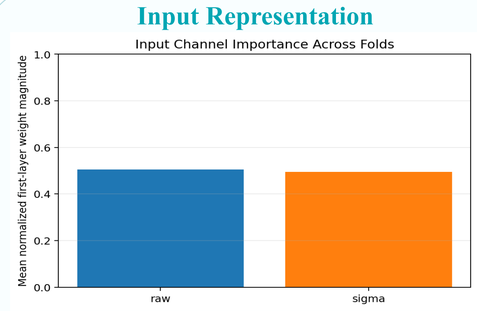
Qualitative Error Analysis



The example shows true detection, false-positive spindle-like activity, and boundary mismatch. Most errors are not complete failures, but over-detections or onset/offset localization errors.

True Positive | **False Positive** | **Boundary Error**

Input Representation



Raw EEG and sigma-band EEG contributed almost equally, supporting the dual-input design.

Reproducibility

- Python 3.11, PyTorch 2.6 + CUDA 12.4.
- Saved fold splits, configurations, checkpoints, probabilities, predicted events, and paper-ready outputs.
- Subject-wise split prevents subject leakage.

INTERPRETATION AND TAKE-HOME MESSAGE

What the Model Learns

- Dense segmentation directly predicts sample-wise spindle probabilities.
- Attention-gated skip connections preserve temporal localization.
- Raw + sigma inputs capture complementary morphology and oscillatory power.
- Ensembling and calibration stabilize event-level decoding.

Why External Validation Matters

MASS-SS2 strict F1: 0.805

DREAMS zero-shot strict F1: 0.476

Approx. 41% relative reduction

This decline is scientifically important: it reveals real cross-dataset variability that internal validation alone cannot detect.

Take-Home Message

- The framework achieves stable MASS-SS2 performance with strict F1 = 0.805.
- Annotation variability is a major factor: F1 changes from 0.563 (E1) to 0.805 (union).
- Zero-shot DREAMS validation exposes generalization limits and motivates domain-aware training.
- The pipeline provides a reproducible benchmark for annotation-aware spindle segmentation.

Funding: H2TRAIN project is supported by the Chips Joint Undertaking (GA 101140052) and its members, including top-up funding by the National Funding Agency for Finland, the Innovation Funding Agency Business Finland.

***Contact*:** waqas.ahmed@uef.fi

BUSINESS FINLAND

Chips JU

H2TRAIN