6 (2026) 100295

Contents lists available at ScienceDirect

Neuroscience Informatics

journal homepage: www.elsevier.com/locate/neuri

Original article

Spindle-AttnUNet: An annotation-aware dual-input attention-gated 1D U-Net for subject-independent N2 sleep spindle segmentation

Waqas Ahmed *, Keijo Haataja, Pekka Toivanen

School of Computing, University of Eastern Finland, Kuopio, Finland

ARTICLE INFO

Keywords:
Sleep spindle detection
EEG
Deep learning
Attention-gated U-Net
Sleep analysis
Annotation variability
External validation
Cross-dataset generalization

ABSTRACT

Background: Sleep spindles are important biomarkers of sleep physiology and neurological function. Automated spindle detection remains challenging due to inter-subject variability, class imbalance, inconsistent expert annotations, and limited cross-dataset generalization.

New Method: We propose an attention-gated dual-input one-dimensional U-Net for subject-independent sleep spindle detection from single-channel EEG. The framework formulates spindle detection as a dense temporal segmentation task by jointly learning from raw EEG and sigma-band representations, while incorporating imbalance-aware optimization and annotation-aware evaluation.

Results: Experiments on the MASS-SS2 dataset using subject-wise five-fold cross-validation achieved a strict event-level F1-score of 0.805 ± 0.031, a tolerant event-level F1-score of 0.846 ± 0.017, and a sample-level ROC-AUC of 0.933 ± 0.015. External zero-shot validation on the DREAMS dataset yielded a strict event-level F1-score of 0.476 and a sample-level ROC-AUC of 0.873 without retraining.

Comparison with Existing Methods: The proposed framework achieved competitive performance relative to recent deep learning methods while providing additional capabilities, including dual-input feature learning, annotation-aware evaluation across multiple expert references, and external cross-dataset validation, which are rarely investigated simultaneously in existing spindle detection studies.

Conclusions: The proposed framework provides robust and reproducible sleep spindle detection with competitive performance and transferable spindle representations. The findings highlight the importance of annotation-aware benchmarking and external validation for developing reliable AI-based sleep analysis systems.

1. Introduction

Sleep spindles are transient oscillatory events observed in the electroencephalogram (EEG), typically occurring between the sigma frequency band 11 and 16 Hz during stage N2 of non-rapid eye movement (NREM) sleep [1,2]. These rhythmic bursts, lasting between 0.5 and 3 s, are generated through thalamocortical interactions and reflect coordinated neural activity between the thalamus and cortex [3]. Beyond their electrophysiological signature, sleep spindles are increasingly recognized as markers of synaptic plasticity and sleep-dependent memory consolidation [4,5], contributing to cognitive processes such as learning and information integration [6,7].

From a clinical perspective, alterations in spindle characteristics such as density, amplitude, and frequency have been associated with a range of neurological and psychiatric conditions, including schizophrenia [8,9], Alzheimer's disease [6,10], and developmental disorders [10,11]. Consequently, reliable quantification of spindle activity is of considerable interest for both research and diagnostic purposes.

However, manual annotation remains labor-intensive and subject to inter-rater variability, limiting scalability and reproducibility [12]. Accurate and automated detection methods are therefore essential to enable large-scale studies, improve diagnostic consistency, and facilitate the integration of spindle-related biomarkers into clinical practice [13].

Although sleep spindles are commonly studied as physiological and clinical biomarkers, accurate event detection does not automatically imply that automatically derived spindle characteristics are clinically valid. A detector may achieve high agreement with expert event annotations while still introducing systematic differences in spindle density, duration, amplitude, or frequency. Therefore, in addition to event-level and sample-level detection metrics, it is important to assess whether model-derived spindle characteristics are consistent with those obtained from manual annotations. In the present study, we evaluate agreement

* Corresponding author.
E-mail addresses: waqas.ahmed@uef.fi (W. Ahmed), keijo.haataja@uef.fi (K. Haataja), pekka.toivanen@uef.fi (P. Toivanen).

https://doi.org/10.1016/j.neuri.2026.100295
Received 5 July 2026; Received in revised form 28 August 2026; Accepted 4 September 2026
Available online 11 September 2026

with expert annotations and basic spindle-event characteristics; however, validation against cognitive outcomes, age-related changes, diagnosis, medication status, or disease severity remains outside the scope of the available data.

1.1. Research aim and contributions

This study aims to develop, validate, and externally assess a subject-independent deep learning framework for automatic sleep spindle detection from single-channel EEG. The proposed approach formulates spindle detection as a dense temporal segmentation problem using an attention-gated 1D U-Net and investigates the effects of expert annotation variability on detection performance. In addition, the transferability of the learned spindle representations is assessed through external validation on the DREAMS dataset.

The main contributions of this work are as follows:

- A novel attention-gated 1D U-Net framework for dense sleep spindle segmentation using complementary raw EEG and sigma-band EEG representations.
- A robust training and inference pipeline incorporating BCE-FPFN optimization, probability calibration, event-level post-processing, and ensemble prediction to address class imbalance and improve detection stability.
- A comprehensive expert-aware evaluation against Expert 1, Expert 2, and union annotations, providing quantitative insights into the impact of annotation variability on reported spindle detection performance.
- Extensive validation through subject-wise five-fold cross-validation on MASS-SS2 and strict zero-shot external evaluation on the DREAMS dataset, demonstrating both robustness and cross-dataset generalization.
- A fully reproducible experimental framework that supports transparent benchmarking and future comparison of automated spindle detection methods.

The proposed framework is intended as a methodological tool for reproducible automated N2 sleep spindle segmentation and quantitative sleep EEG analysis. It is not proposed as a validated clinical diagnostic system. Clinical deployment would require additional validation across heterogeneous patient populations, recording systems, EEG references, and clinically relevant outcomes.

1.2. Paper structure

The remainder of this paper is organized as follows. Section 2 reviews related work on automated sleep spindle detection. Section 3 presents the proposed methodology, including dataset preparation, model architecture, training strategy, and evaluation protocol. Section 4 reports the experimental results, ablation analysis, expert-specific evaluation, and external validation on the DREAMS dataset. Finally, Sections 5 and 6 discuss the findings and conclude the study.

2. Literature review

Automatic sleep spindle detection has been extensively studied using signal processing, machine learning, and deep learning approaches. Early methods primarily relied on handcrafted features derived from EEG signals, such as spectral power and envelope-based thresholding. Traditional machine learning techniques improved detection performance by incorporating feature-based classification. More recently, deep learning methods have achieved state-of-the-art performance by learning hierarchical representations directly from raw EEG signals.

Ramirez et al. [14] proposed an unsupervised sleep spindle detection framework, named USSD, based on dictionary learning and adaptive thresholding. The method was evaluated on MASS-SS2, INTA-UCH, and CAP datasets using event-based recall, precision, and F1-score with IoU ≥ 0.2, achieving an F1-score of approximately 0.72, which improved to 0.78 after fine-tuning with 20% labeled MASS-SS2 data. Instead of handcrafted features, USSD learns spindle-like dictionary atoms of different durations and applies convolution-based matching, AASM-based post-processing, and probabilistic event scoring.

Zapata et al. [15] proposed a multitaper-based sleep spindle detection framework (SAMC) combining spectral estimation and convolution for multichannel EEG analysis. The method was evaluated on MASS-SS2, NAP, and DREAMS datasets using agreement rate (AR), sensitivity, positive predictive value (PPV), F-score, and Lin's concordance correlation coefficient (CCC), achieving on MASS-SS2 an AR of 96%, sensitivity of 0.96, PPV of 0.92, and F-score of 0.94.

Kaulen et al. [16] proposed a deep learning–based sleep spindle detection framework using a U-Net-type architecture (SUMO) trained on expert-consensus annotations from the MODA dataset. The architecture consists of a fully convolutional encoder–decoder with skip connections that capture temporal context at multiple scales. Evaluated on a held-out test set, the method achieved an F1-score of 0.82, outperforming both the A7 algorithm (0.73) and average human experts (0.72), with precision 0.85 and recall 0.79.

Tapia-Rivas et al. [17] introduced the Sleep EEG Event Detector (SEED), a deep learning framework for detecting sleep spindles (SSs) and K-complexes (KCs) from single-channel EEG. SEED combines convolutional neural networks for local feature extraction with bidirectional LSTM layers to model long-range temporal context (approximately 20 s), enabling precise detection and localization of transient sleep events. Evaluated on the MASS2 dataset, SEED achieved state-of-the-art performance, with F1-scores of approximately 80.5% for sleep spindles and 83.7% for K-complexes.

Jiang et al. [18] proposed a robust two-stage sleep spindle detection framework based on single-channel EEG. Evaluated on the MASS and DREAMS datasets, the method achieved competitive performance, reaching an F1-score of approximately 81.4% on MASS and 69% on DREAMS under event-based evaluation, demonstrating strong generalization and robustness to inter-subject variability.

Pan et al. [19] proposed a deep learning-augmented sleep spindle detection framework (CDTSD) tailored for both healthy subjects and patients with acute disorders of consciousness (ADOC). The method adopts a hybrid two-stage architecture combining a convolutional neural network (CNN) with a decision tree-based validation mechanism. The method achieves strong performance, with F1-scores of 79.8% and 84.1% on the MASS dataset and 74.5% on an ADOC dataset, demonstrating robustness across heterogeneous EEG conditions and improved detection of clinically relevant spindle patterns.

Rychlík et al. [20] investigated automated sleep spindle detection using machine and deep learning approaches on the MASS SS2 dataset. The study employed multiple neural architectures. Feature learning was performed directly from EEG segments, treating signals as structured inputs for temporal and spatial pattern extraction. Evaluation was conducted using classification accuracy, where the CNN-LSTM (Torch implementation) achieved the best performance of 67.15%, followed by DNN (64.63%) and CNN (60.87%).

Wei et al. [21] proposed Deep-Spindle, a deep learning-based (CNN and LSTM) automated sleep spindle detection system using infant EEG data. The model was trained and evaluated on infant EEG datasets (ex-term and ex-preterm), with feature learning performed implicitly without handcrafted extraction. Performance evaluation included sensitivity, specificity, precision, F1-score, accuracy, and Matthews correlation coefficient (MCC), achieving 91.9%–96.5% sensitivity, 95.3%–96.7% specificity, and F1-score up to 0.954 on independent test sets.

Yang et al. [22] proposed SpindleCatcher, a deep learning (CNN) framework for automatic sleep spindle detection with application to acute disorders of consciousness (ADOC). Experiments were conducted on the MASS SS2 dataset and a clinical ADOC dataset, without explicit handcrafted feature engineering, using recall and F1-score as evalua-

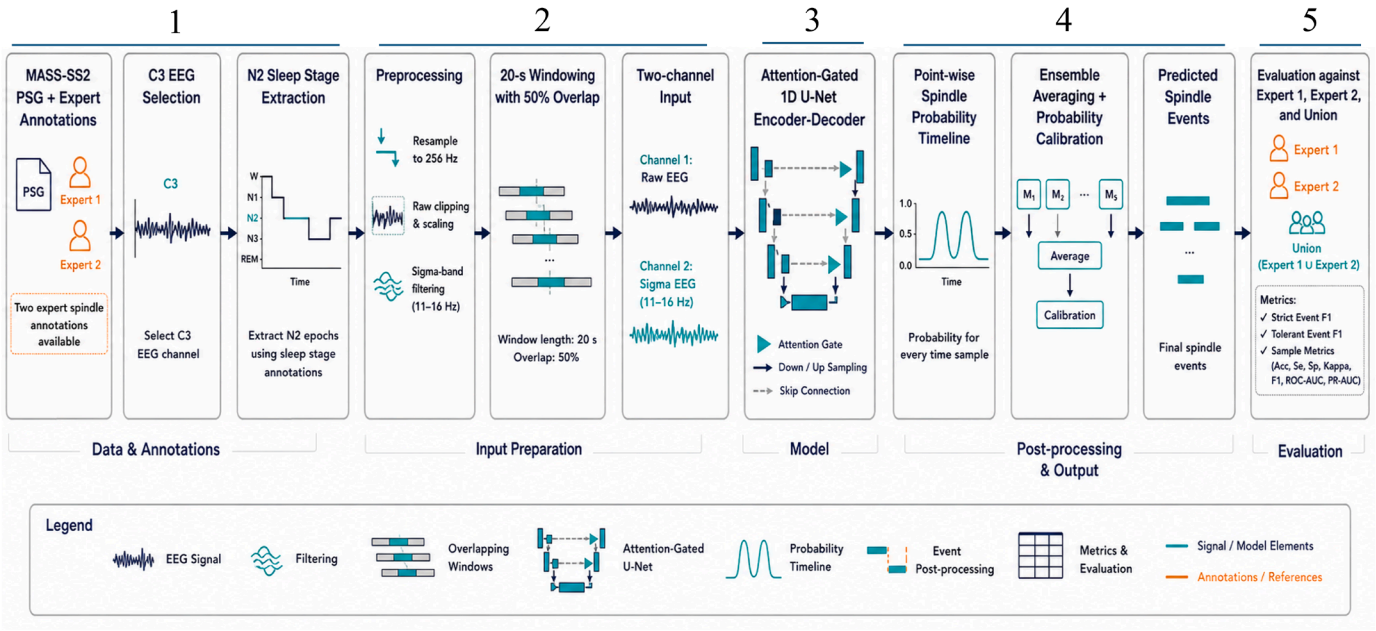

Fig. 1. Overview of the proposed attention-gated 1D U-Net framework for N2 sleep spindle detection using raw EEG and sigma-band EEG representations.

tion metrics. The method achieved a recall of (81.7%) and F1-score of (79.4%) on MASS SS2.

You et al. [23] introduced SpindleU-Net, a deep learning framework based on a one-dimensional U-Net architecture with an attention mechanism for point-wise sleep spindle detection. The method was evaluated on the MASS SS2 dataset and DREAMS dataset, using raw EEG with normalization and data augmentation, without explicit handcrafted feature extraction. Performance was assessed using precision, recall, and F1-score under event-based and point-wise evaluation, achieving F1-scores of 85.4% (E1) and 80.3% (E2) on MASS SS2.

Tian and Wei [24] proposed a U-Net-based deep learning framework enhanced with a label optimization module using crowdsourced EEG annotations from the MODA dataset and MASS subsets. Performance evaluation included sample-event and subject-level analyses with metrics including F1-score, MCC, Kappa coefficient, precision, and sensitivity. Experiments achieved an F1-score of 82.0% at an overlap threshold of 0.6, along with superior agreement with expert annotations in spindle feature extraction.

Chen et al. [25] proposed a deep neural network-based framework that combines CNN-extracted deep features with entropy-based features for sleep spindle detection using the DREAMS EEG dataset. Performance was evaluated using precision, recall, and F1-score, achieving an F1-score of 66.0%, with improvements observed when incorporating entropy features.

Hu et al. [26] proposed a unified deep learning framework based on a Bidirectional Long Short-Term Memory (BiLSTM) architecture for joint detection of sleep spindles and K-complexes from EEG signals. Evaluated on the DREAMS and MASS (SS2) datasets using event-level metrics (precision, recall, F1-score), the framework achieved state-of-the-art performance, including F1-scores of 79.6% (spindles, DREAMS), 93.8% (K-complexes, DREAMS), 88.2% (spindles, MASS), and 90.3% (K-complexes, MASS), demonstrating significant improvements over existing methods.

Sai Babu et al. [27] proposed an automatic sleep spindle detection framework (SST-SMOTE-Adaboost) that combines Synchrosqueezed Wavelet Transform (SWT) for time–frequency feature extraction, Synthetic Minority Oversampling Technique (SMOTE) for class imbalance handling, and the Adaboost algorithm for classification. Evaluated on the MASS (SS2 subset) dataset using event-based metrics, the approach achieved an F1-score of approximately 76%, sensitivity of 78.9%, and precision (PPV) of 73.2%. The results demonstrate that integrating imbalance-aware sampling with discriminative time–frequency features improves spindle detection performance while maintaining computational simplicity.

Kaulen et al. [28] proposed SUMO (Slim U-Net trained on MODA), a deep learning framework based on a lightweight U-Net architecture. Evaluated using event-based metrics with a 20% overlap threshold on MODA dataset. The model achieved an F1-score of approximately 82%. Additionally, SUMO demonstrated strong generalization across age groups, achieving F1-scores of 84% for younger and 79% for older subjects.

Gurdiel et al. [29] proposed SpinCo, an interpretable machine learning framework for sleep spindle detection based on exhaustive feature extraction and the XGBoost algorithm. Evaluated on MASS, DREAMS, and a pediatric clinical datasets. The model demonstrated strong agreement with expert annotations ($R^2 > 0.6$ for spindle characteristics) and exhibited inter-expert generalization comparable to human agreement levels, indicating near human-level performance.

Overall, existing techniques present a trade-off between performance and interpretability. Deep learning models achieve high accuracy but lack transparency, whereas traditional machine learning methods offer interpretability but may underperform in complex scenarios.

3. Research methodology

3.1. Overview of the proposed framework

This study proposes a subject-independent sleep spindle detection framework based on an attention-gated one-dimensional U-Net. As shown in Fig. 1, the framework includes data & annotation, input preparation, U-Net model, post-processing & output, and evaluation against Expert 1, Expert 2, and union reference annotations.

The problem is formulated as dense temporal segmentation. Each EEG sample is assigned a spindle probability, and the resulting probability sequence is converted into discrete spindle events using validation-based post-processing. The final model was trained using union annotations and evaluated using both sample-level and event-level metrics.

Let $X_i \in \mathbb{R}^{C \times T}$ denote the ith input window, where C is the number of input channels and T is the number of temporal samples. The model

Table 1
Subject-wise MASS-SS2 summary after N2-stage restriction. The union reference is an inclusive composite annotation containing events marked by either expert.

Subject	E1 Count	E2 Count	Union Count	E1 Median (s)	E2 Median (s)	Union Median (s)
1	1068	2531	2445	0.793	1.098	1.129
2	1174	2292	2223	0.801	1.137	1.188
3	145	613	608	0.648	0.891	0.896
5	346	1227	1204	0.684	0.859	0.867
6	150	862	847	0.750	0.803	0.832
7	933	1660	1671	0.852	1.188	1.219
9	835	1729	1683	0.875	1.211	1.262
10	813	2004	1953	0.773	1.084	1.117
11	618	1592	1546	0.875	1.137	1.188
12	719	1246	1243	0.809	1.090	1.125
13	712	1495	1464	0.855	1.410	1.455
14	724	1670	1638	0.773	1.098	1.137
17	488	1235	1204	0.773	1.238	1.250
18	1188	1722	1786	0.824	0.914	0.957
19	321	1091	1061	0.742	0.949	0.980

estimates a sample-wise spindle probability sequence as

$$\hat{\mathbf{p}}_i = f_\theta(X_i), \tag{1}$$

where f_θ denotes the attention-gated 1D U-Net and $\hat{\mathbf{p}}_i \in [0,1]^T$.

To assess the external generalization capability of the proposed model, we additionally evaluated it on the DREAMS sleep spindle dataset [30]. DREAMS contains eight 30-min PSG excerpts recorded from subjects with different sleep-related disorders, with original sampling frequencies ranging from 50 to 200 Hz. Each excerpt was annotated for sleep spindles by expert scorers; one expert annotated all eight excerpts, while the second expert annotated six excerpts. Consistent with the available DREAMS derivations, excerpts 1 and 3 were evaluated using the C3-A1 EEG channel, whereas the remaining excerpts were evaluated using the closest central derivation, CZ-A1. All EEG signals were resampled to 256 Hz to match the MASS-SS2 preprocessing pipeline. The MASS-trained model was applied to DREAMS in a zero-shot setting, without retraining or fine-tuning, and the union of the available expert annotations was used as the primary reference standard.

3.2. Dataset, EEG derivation, and annotation strategy

The experiments were conducted on the MASS-SS2 subset of the Montreal Archive of Sleep Studies [31]. We analyzed the central EEG derivation C3–A1, which matched the derivation used for spindle annotation. A subject was retained only when the polysomnography recording, sleep-stage labels, and both expert spindle annotation files were available. This resulted in 15 subjects.

Each retained recording included spindle annotations from two independent experts, denoted as Expert 1 (E1) and Expert 2 (E2). Table 1 summarizes the subject-wise spindle counts and median durations after N2-stage restriction. The annotations showed clear inter-scorer variability, especially in event counts, with E2 generally marking more spindles than E1.

For training, we used an inclusive union-based reference. For subject s and sample index t, the label was defined as

$$y_s^{\text{union}}(t) = \mathbb{I}\left(y_s^{E1}(t) = 1 \vee y_s^{E2}(t) = 1\right), \tag{2}$$

where $y_s^{E1}(t)$ and $y_s^{E2}(t)$ are the binary annotations from E1 and E2. At the event level, this corresponds to

$$E_{\text{union}} = E_1 \cup E_2. \tag{3}$$

The union was used as an inclusive composite reference, not as expert consensus. It reduces the risk of treating plausible single-expert spindles as negatives, but it may also reflect the more liberal scorer. Therefore, final performance was reported separately against E1, E2, and the union reference.

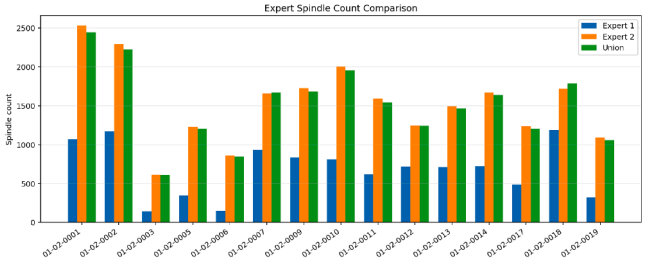

Fig. 2. Subject-wise spindle-count comparison for Expert 1, Expert 2, and the union-based inclusive reference.

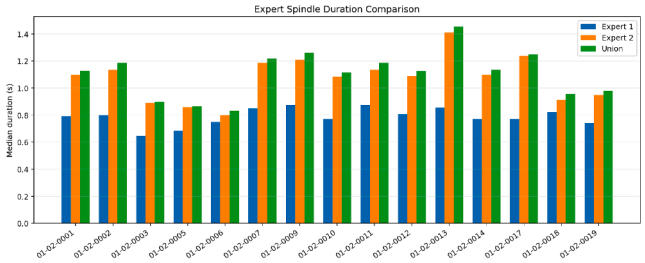

Fig. 3. Subject-wise median-duration comparison for Expert 1, Expert 2, and the union-based inclusive reference.

To examine annotation uncertainty, we also considered an agreement-aware soft-label formulation in the ablation analysis:

$$y_t^{\text{soft}} = \begin{cases} 1, & t \in E_1 \cap E_2, \\ 0.5, & t \in (E_1 \cup E_2) \setminus (E_1 \cap E_2), \\ 0, & t \notin E_1 \cup E_2. \end{cases} \tag{4}$$

This formulation preserves inter-expert disagreement instead of converting all annotated samples into the same positive class.

Fig. 2 shows the subject-wise spindle counts for E1, E2, and the union reference. Fig. 3 compares the corresponding median durations. These summaries highlight both count-level and boundary-level annotation variability.

3.3. EEG preprocessing

Sleep-stage annotations were parsed to retain only N2 sleep. Let $x(t)$ denote the C3–A1 EEG signal and Ω_{N2} the set of N2 samples. The retained signal was defined as

$$x_{\text{N2}}(t) = x(t), \quad t \in \Omega_{\text{N2}}. \tag{5}$$

All signals were resampled to $f_s = 256$ Hz. The raw EEG was converted to microvolts when required, clipped to ± 150 μV, and scaled as

$$x_{\text{raw}}(t) = \frac{\text{clip}(x(t), -A, A)}{2A}, \quad A = 150\ \mu\text{V}. \tag{6}$$

This step limited extreme amplitudes while preserving spindle morphology.

A sigma-band representation was extracted using a 12–16 Hz bandpass filter:

$$x_\sigma(t) = \mathcal{F}^{-1}\left[H_\sigma(f)\mathcal{F}\{x_{\text{raw}}(t)\}\right], \tag{7}$$

where $H_\sigma(f)$ denotes the sigma-band filter. The final model input was

$$X_i = [x_{\text{raw}}, x_\sigma] \in \mathbb{R}^{2\times T}. \tag{8}$$

3.4. Segmentation window construction

The retained N2 timeline was divided into 20-s windows. With $f_s = 256$ Hz, each window contained

$$T = 20 \times 256 = 5120 \tag{9}$$

samples. A 50% overlap was used between adjacent windows. The hop length was therefore

$$H = T(1-\rho), \tag{10}$$

where $\rho = 0.5$. Each window had a binary target sequence

$$Y_i = [y_1, y_2, \ldots, y_T], \tag{11}$$

where $y_t = 1$ indicates that sample t belongs to a union-annotated spindle.

3.5. Attention-gated 1D U-Net architecture

The proposed model uses an encoder-decoder 1D U-Net architecture. The encoder extracts hierarchical temporal features using convolutional blocks and max-pooling. The decoder restores the original temporal resolution using transposed convolutions and skip connections.

Each convolutional block consists of stacked one-dimensional convolutions, batch normalization, ReLU activation, and dropout. A generic convolutional transformation at layer l is

$$\mathbf{h}^{(l)} = \phi\big(\text{BN}\big(\mathbf{W}^{(l)} * \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\big)\big), \tag{12}$$

where $*$ denotes 1D convolution, $\text{BN}(\cdot)$ is batch normalization, and $\phi(\cdot)$ is the ReLU function.

Attention gates were applied to encoder skip features before fusion with decoder features. Let $\mathbf{s}$ denote the encoder skip feature and $\mathbf{g}$ denote the decoder gating feature. The attention coefficient is computed as

$$\alpha = \sigma\left[\psi^T\phi\big(W_s\mathbf{s} + W_g\mathbf{g}\big)\right], \tag{13}$$

and the gated skip feature is

$$\tilde{\mathbf{s}} = \alpha \odot \mathbf{s}, \tag{14}$$

where $\sigma(\cdot)$ is the sigmoid function and $\odot$ denotes element-wise multiplication. The output layer produces a sample-wise logit sequence z_t, which is converted to a spindle probability by

$$\hat{p}_t = \sigma(z_t). \tag{15}$$

3.6. Loss function and optimization

The model was trained using a combined binary cross-entropy and false-positive/false-negative penalty loss. The binary cross-entropy term was

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{T}\sum_{t=1}^{T}\left[w y_t \log(\hat{p}_t) + (1-y_t)\log(1-\hat{p}_t)\right], \tag{16}$$

where w is the positive-class weight.

The false-positive/false-negative penalty was defined as

$$\mathcal{L}_{\text{FPFN}} = \frac{FP}{N_{\text{neg}} + \epsilon} + \frac{FN}{N_{\text{pos}} + \epsilon}, \tag{17}$$

where FP and FN denote soft false-positive and false-negative counts, N_{neg} and N_{pos} are the numbers of negative and positive samples, and ϵ is a small numerical constant.

The final training loss was

$$\mathcal{L} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{FPFN}}\mathcal{L}_{\text{FPFN}}. \tag{18}$$

In the final configuration, $\lambda_{\text{BCE}} = 1.0$ and $\lambda_{\text{FPFN}} = 1.0$. Positive-window sampling was enabled to address class imbalance. Windows containing spindle activity were assigned a sampling weight of 3.0. The model was optimized using Adam with a learning rate of 10^{-4}, learning-rate scheduling, mixed-precision training, and early stopping.

3.7. Ensemble probability estimation

To improve robustness, five U-Net models were trained using different random seed offsets. During inference, each model produced a probability timeline for the same EEG input. The final probability was obtained by averaging the ensemble predictions:

$$\hat{p}_t = \frac{1}{M}\sum_{m=1}^{M} f_{\theta_m}(X)_t, \tag{19}$$

where $M = 5$ and f_{θ_m} denotes the mth trained model.

Probability calibration was applied using validation subjects. Platt scaling was used to map uncalibrated probabilities to calibrated probabilities:

$$\tilde{p}_t = \frac{1}{1 + \exp[-(a\hat{p}_t + b)]}, \tag{20}$$

where a and b were estimated from validation data.

3.8. Inference and event-level post-processing

During inference, each test subject's N2 EEG was processed using overlapping 20-s windows. Window predictions were stitched to form a continuous subject-level probability timeline. Overlapping predictions were averaged.

The calibrated probability sequence was binarized using a validation-selected threshold τ:

$$\hat{y}_t = \begin{cases} 1, & \tilde{p}_t \geq \tau, \\ 0, & \tilde{p}_t < \tau. \end{cases} \tag{21}$$

Consecutive positive samples were grouped into predicted spindle events. Events shorter than the selected minimum duration were removed. In the final search space, the probability threshold was selected from

$$\tau \in \{0.55, 0.575, 0.60, 0.625, 0.65, 0.675, 0.70\}, \tag{22}$$

the merge gap from

$$g \in \{0.05, 0.10, 0.15\}\ \text{s}, \tag{23}$$

and the minimum event duration was fixed at

$$d_{\min} = 0.6\ \text{s}. \tag{24}$$

Adaptive threshold quantiles were selected from

$$q \in \{0.85, 0.90, 0.95, 1.0\}. \tag{25}$$

All threshold and post-processing parameters were selected using validation subjects only.

3.9. Density-penalized validation selection

The validation objective combined mean F1, minimum subject-level F1, and mean precision. Let $\bar{F}_1$ denote mean validation F1, $F_{1,\min}$ denote the worst-subject F1, and $\bar{P}$ denote mean precision. The base selection score was

$$J_{\text{base}} = 0.75\bar{F}_1 + 0.15F_{1,\min} + 0.10\bar{P}. \tag{26}$$

A density penalty was used to avoid over-detection or under-detection relative to the reference spindle density. Let r denote the ratio between predicted and reference event density. The density penalty was

$$\Pi(r) = \begin{cases} 0.20(r - 1.10), & r > 1.10, \\ 0.10(0.80 - r), & r < 0.80, \\ 0, & 0.80 \le r \le 1.10. \end{cases} \tag{27}$$

The final validation score was

$$J = J_{\text{base}} - \Pi(r). \tag{28}$$

The parameter set with the highest J was selected within each fold.

3.10. Cross-validation protocol

A subject-wise five-fold cross-validation protocol was used. In fold k, the retained subjects were divided into disjoint training, validation, and test sets:

$$S = S_{\text{train}}^{(k)} \cup S_{\text{val}}^{(k)} \cup S_{\text{test}}^{(k)}, \tag{29}$$

with

$$S_{\text{train}}^{(k)} \cap S_{\text{val}}^{(k)} = S_{\text{train}}^{(k)} \cap S_{\text{test}}^{(k)} = S_{\text{val}}^{(k)} \cap S_{\text{test}}^{(k)} = \emptyset. \tag{30}$$

Training subjects were used for parameter optimization. Validation subjects were used for early stopping, calibration, threshold selection, and post-processing selection. Test subjects were used only for final evaluation. This design prevents subject-level data leakage.

3.11. Evaluation metrics

The proposed framework was evaluated using sample-level and event-level metrics.

Sample-level precision, recall, and F1-score were computed as

$$P = \frac{TP}{TP + FP}, \tag{31}$$

$$R = \frac{TP}{TP + FN}, \tag{32}$$

$$F1 = \frac{2PR}{P + R}. \tag{33}$$

Cohen's kappa was computed as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{34}$$

where p_o is the observed agreement and p_e is the agreement expected by chance.

For event-level evaluation, predicted events were matched to reference events using temporal overlap. The intersection-over-union between a predicted event D and an annotated event A was

$$\text{IoU}(D, A) = \frac{|D \cap A|}{|D \cup A|}. \tag{35}$$

Strict event metrics were computed using overlap-based matching. Tolerant event metrics were computed using a temporal tolerance window to account for small boundary discrepancies between predicted and annotated spindles.

Table 2
Main model configuration and training parameters used in the proposed segmentation framework.

Parameter	Value
Model mode	segmentation_unet
Input channel	C3
Annotation mode	union
Stage mode	n2_only
Sampling frequency	256 Hz
Segmentation window	20 s
Window overlap	50%
Input channels	Raw EEG, Sigma EEG
U-Net base channels	32
U-Net depth	4
Kernel size	11
Attention gates	Enabled
Temporal bottleneck	Disabled
Auxiliary heads	Disabled
Loss function	BCE-FPFN loss
Learning rate	1×10^{-4}
Batch size	32
Maximum epochs	200
Early stopping patience	20
Ensemble seed offsets	{0,1,2,3,4}
Threshold grid	{0.55, 0.575, 0.60, 0.625, 0.65, 0.675, 0.70}
Merge-gap grid (s)	{0.05, 0.10, 0.15}
Minimum duration grid (s)	{0.6}
Density-adaptive thresholding	Disabled
Candidate consensus filtering	Disabled

All event-level metrics were reported with respect to E1, E2, and union annotations. Final results were summarized as mean and standard deviation across folds:

$$\bar{m} = \frac{1}{K}\sum_{k=1}^{K} m_k, \tag{36}$$

$$\sigma_m = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(m_k - \bar{m})^2}. \tag{37}$$

3.12. Reproducibility

The proposed framework was implemented in Python using the PyTorch deep learning library. To support reproducible experimentation, the complete source code, final Python pipeline, experiment configurations, fold assignments, evaluation outputs, paper-ready tables, and figures are publicly available in the GitHub repository: https://github.com/waqas43u/attention-gated-spindle-unet.git. During execution, the pipeline automatically archives model hyperparameters, training settings, checkpoints, probability outputs, event predictions, and fold-wise performance summaries.

Table 2 summarizes the principal architectural, training, and post-processing parameters used throughout the study. The final configuration employed an attention-gated 1D U-Net operating on two input channels (raw EEG and sigma-band EEG), trained using union annotations under a subject-independent five-fold cross-validation protocol. Event-level decoding parameters, including threshold selection, merge-gap constraints, and minimum spindle duration, were optimized exclusively on validation data to avoid information leakage.

4. Results interpretations

4.1. Overall detection performance

Table 3 summarizes the overall five-fold performance of the proposed attention-gated 1D U-Net framework under the union annotation strategy. The proposed model achieved a strict event-level F1-score of 0.805 ± 0.031, with precision and recall of 0.812 ± 0.050 and

Table 3
Overall five-fold performance of the proposed attention-gated 1D U-Net under union annotation.

Metric	Mean ± SD
Strict event precision	0.812 ± 0.050
Strict event recall	0.802 ± 0.047
Strict event F1	0.805 ± 0.031
Tolerant event precision	0.853 ± 0.055
Tolerant event recall	0.842 ± 0.029
Tolerant event F1	0.846 ± 0.017
Accuracy (ACC)	0.892 ± 0.023
Sensitivity (SEN)	0.752 ± 0.037
Specificity (SPE)	0.911 ± 0.026
Cohen's κ	0.575 ± 0.037
Sample F1	0.638 ± 0.031
Sample ROC-AUC	0.933 ± 0.015
Sample PR-AUC	0.684 ± 0.060

0.802 ± 0.047, respectively. The close agreement between precision and recall indicates a balanced detection behavior with neither excessive false positives nor excessive missed spindle events.

Under the tolerant evaluation criterion, the model achieved a precision of 0.853 ± 0.055, a recall of 0.842 ± 0.029, and an F1-score of 0.846 ± 0.017. The improvement relative to strict evaluation suggests that many prediction errors arise from minor temporal boundary discrepancies rather than complete event misses. This observation is consistent with the inherent uncertainty of manual spindle onset and offset annotations.

At the sample level, the framework achieved an F1-score of 0.638 ± 0.031, a ROC-AUC of 0.933 ± 0.015, and a PR-AUC of 0.684 ± 0.060. The high ROC-AUC demonstrates strong discrimination between spindle and non-spindle samples, whereas the lower PR-AUC reflects the substantial class imbalance characteristic of sleep spindle detection tasks.

The model further achieved an overall accuracy of 0.892 ± 0.023, sensitivity of 0.752 ± 0.037, and specificity of 0.911 ± 0.026. The higher specificity indicates effective suppression of false spindle detections while maintaining satisfactory sensitivity to true spindle events. In addition, Cohen's κ reached 0.575 ± 0.037, indicating moderate-to-substantial agreement between automated detections and expert annotations beyond chance.

Finally, the validation-selected operating points corresponded to a target spindle density of 6.621 ± 0.811 events per minute, suggesting that the proposed post-processing and threshold optimization procedures produced physiologically plausible spindle densities while preserving event-level detection performance across folds.

4.2. Fold-wise performance

Fig. 4 and Table 4 summarize the fold-wise performance of the proposed framework under subject-independent five-fold cross-validation. The results demonstrate consistent performance across folds despite substantial inter-subject variability in spindle density and annotation characteristics.

The strict event-level F1-score ranged from 0.755 in Fold 1 to 0.833 in Fold 3, with Folds 2–4 exhibiting highly consistent performance above 0.82. Fold 3 achieved the best overall event-level performance, attaining the highest strict precision (0.858), highest Cohen's κ (0.630), and highest sample-level F1-score (0.677). These results indicate that the proposed framework generalized effectively across most subject partitions while maintaining balanced precision and recall.

Fold 1 exhibited the lowest strict event F1-score due primarily to reduced recall (0.715), indicating that a larger proportion of spindle events remained undetected in this fold. In contrast, Fold 5 achieved the highest recall (0.859) but the lowest precision (0.720), suggesting a more liberal detection behavior that increased sensitivity at the expense of additional false positives. The resulting reduction in strict event F1 demonstrates the expected precision–recall trade-off associated with subject-independent spindle detection.

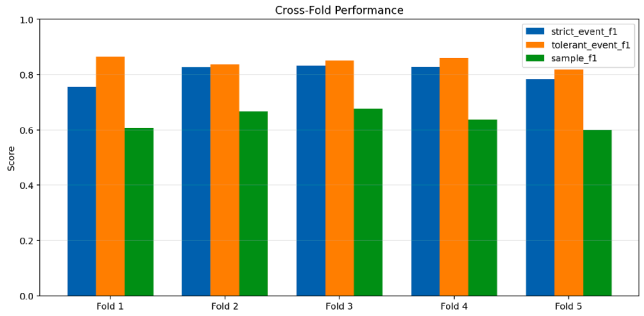

Fig. 4. Cross-fold performance of the proposed model using strict event F1, tolerant event F1, and sample F1.

The tolerant event F1-score remained comparatively stable across folds, varying only from 0.818 to 0.864. The smaller variation relative to strict event F1 indicates that most fold-specific errors originated from inaccuracies in spindle boundary localization rather than complete event omissions. This observation is consistent with the inherent ambiguity of spindle onset and offset annotation and supports the robustness of the proposed segmentation-based framework.

Sample-level metrics further confirm stable discrimination performance across folds. Sample PR-AUC ranged from 0.600 to 0.755, while Cohen's κ varied between 0.535 and 0.630. The relatively narrow spread of these metrics suggests that the ensemble attention-gated U-Net maintained consistent sample-level classification capability despite differences in spindle prevalence and annotation complexity among test subjects.

Overall, the fold-wise analysis demonstrates that the proposed framework achieves reliable generalization across independent subject partitions without evidence of severe fold-specific performance degradation. The limited variability observed in both event-level and sample-level metrics indicates that the model learned spindle representations that are robust to inter-subtocol and inter-subject variability within the MASS-SS2 dataset.

4.3. Expert-specific evaluation

To quantify the effect of annotation variability, the proposed framework was evaluated separately against Expert 1 (E1), Expert 2 (E2), and the union-based inclusive reference. Table 5 reports event-level and sample-level performance across the five cross-validation folds.

The detector showed substantially different agreement patterns across the two expert references. Under strict event-level evaluation, the F1-score was 0.805 ± 0.031 for the union-based reference and 0.800 ± 0.029 for E2, but decreased to 0.563 ± 0.051 for E1. This reduction was driven mainly by lower precision against E1 (0.407 ± 0.051), while recall remained high (0.920 ± 0.018).

This pattern reflects the annotation characteristics of the two experts. E2 marked more spindle events than E1, and the union reference therefore followed a more inclusive annotation profile. As a result, many events accepted by E2 or by the union reference were counted as false positives when E1 was used as the reference. The average number of false positives was therefore much higher for E1 (2758 ± 570) than for E2 (938 ± 302) or the union reference (879 ± 302).

The tolerant event-level results showed the same trend. Tolerant F1 reached 0.846 ± 0.017 for the union reference and 0.841 ± 0.018 for E2, but only 0.583 ± 0.050 for E1. The limited improvement for E1 from strict to tolerant scoring suggests that the main discrepancy was not boundary localization, but the different inclusion criteria used by the annotators.

Table 4
Fold-wise performance of the proposed attention-gated 1D U-Net under union annotation.

Fold	Strict Event F1	Strict Precision	Strict Recall	Tolerant Event F1	Sample F1	Sample PR-AUC	Cohen's κ
1	0.755	0.799	0.715	0.864	0.606	0.600	0.556
2	0.826	0.838	0.814	0.837	0.667	0.755	0.608
3	0.833	0.858	0.809	0.850	0.677	0.747	0.630
4	0.827	0.842	0.813	0.860	0.638	0.645	0.549
5	0.784	0.720	0.859	0.818	0.600	0.674	0.535

Table 5
Cross-validation performance against Expert 1, Expert 2, and the union-based inclusive reference. Values are reported as mean ± standard deviation across five folds.

Metric	Expert 1	Expert 2	Union
Strict event precision	0.407 ± 0.051	0.799 ± 0.047	0.812 ± 0.050
Strict event recall	0.920 ± 0.018	0.804 ± 0.048	0.802 ± 0.047
Strict event F1	0.563 ± 0.051	0.800 ± 0.029	0.805 ± 0.031
Strict TP	1886 ± 384	3705 ± 665	3756 ± 684
Strict FP	2758 ± 570	938 ± 302	879 ± 302
Strict FN	161 ± 41	888 ± 183	910 ± 182
Tolerant event precision	0.422 ± 0.051	0.841 ± 0.055	0.853 ± 0.055
Tolerant event recall	0.953 ± 0.007	0.845 ± 0.029	0.842 ± 0.029
Tolerant event F1	0.583 ± 0.050	0.841 ± 0.018	0.846 ± 0.017
Tolerant TP	1949 ± 384	3881 ± 611	3939 ± 624
Tolerant FP	2694 ± 574	763 ± 337	705 ± 331
Tolerant FN	97 ± 25	713 ± 163	736 ± 168
Accuracy	0.862 ± 0.034	0.892 ± 0.025	0.894 ± 0.025
Sensitivity	0.931 ± 0.026	0.754 ± 0.029	0.753 ± 0.029
Specificity	0.859 ± 0.036	0.910 ± 0.027	0.912 ± 0.027
Cohen's κ	0.281 ± 0.056	0.565 ± 0.038	0.572 ± 0.040
Sample F1	0.324 ± 0.058	0.626 ± 0.032	0.639 ± 0.034
Sample ROC-AUC	0.952 ± 0.021	0.934 ± 0.016	0.935 ± 0.016
Sample PR-AUC	0.462 ± 0.090	0.677 ± 0.056	0.686 ± 0.059

Sample-level metrics further supported this interpretation. The sample F1-score was 0.639 ± 0.034 for the union reference and 0.626 ± 0.032 for E2, compared with 0.324 ± 0.058 for E1. Cohen's κ showed a similar trend, increasing from 0.281 ± 0.056 for E1 to 0.572 ± 0.040 for the union reference.

Overall, the model aligned more closely with E2 and with the union-based inclusive reference than with E1. The union reference should therefore be interpreted as an inclusive composite annotation, not as expert consensus. These findings show that reported spindle detection performance depends strongly on the selected annotation reference and support the need for expert-specific reporting.

Fig. 5 and Fig. 6 show the fold-wise strict and tolerant event-level performance for each reference.

4.4. Subject-wise analysis

Subject-wise evaluation revealed substantial variability in detection performance across the 15 retained MASS-SS2 subjects. Table 6 summarizes the strict event-level precision, recall, and F1-score obtained against Expert 1, Expert 2, and union annotations.

Under the union reference standard, strict event-level F1-scores ranged from 0.631 for subject 01-02-0006 to 0.875 for subject 01-02-0010. Additional challenging subjects included 01-02-0002, 01-02-0011, 01-02-0019, and 01-02-0014, all of which achieved union F1-scores below 0.79. In contrast, subjects 01-02-0003, 01-02-0009, 01-02-0010, and 01-02-0017 exhibited strong agreement with the reference annotations, yielding union F1-scores above 0.82.

A consistent pattern was observed across subjects when comparing annotation references. Evaluation against Expert 1 produced markedly lower precision but substantially higher recall than evaluation against Expert 2 or union annotations. For example, subject 01-02-0003 achieved an Expert 1 precision of only 0.222 despite a recall of 0.959,

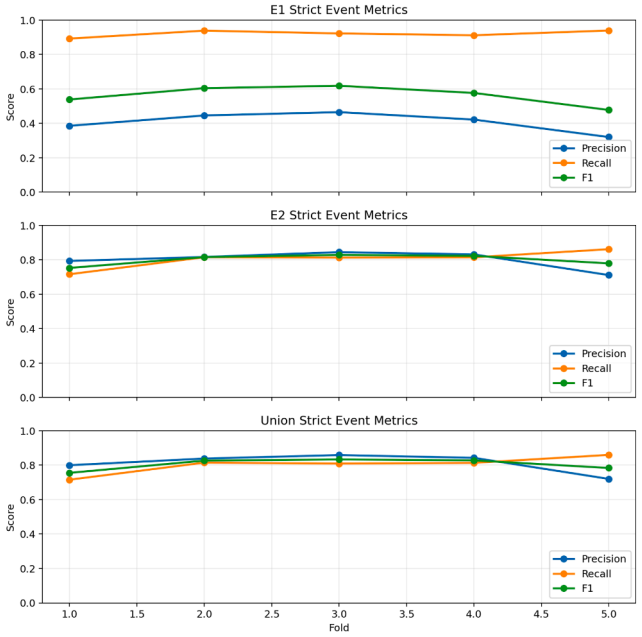

Fig. 5. Fold-wise strict event-level precision, recall, and F1-score against Expert 1, Expert 2, and the union-based inclusive reference.

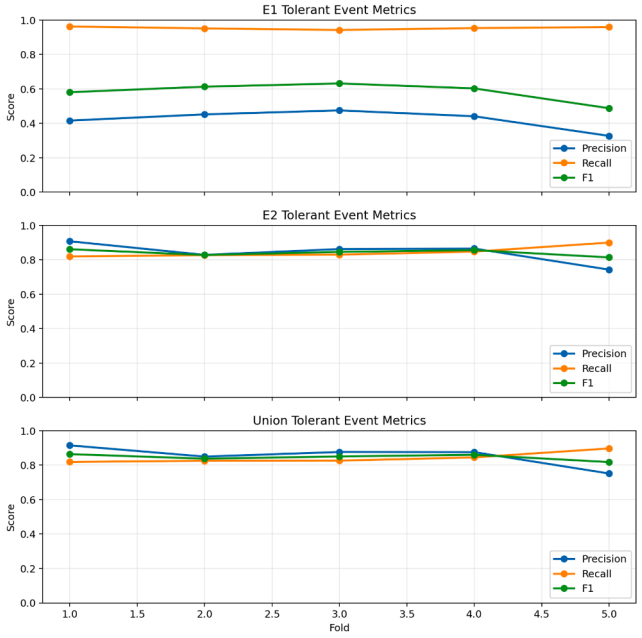

Fig. 6. Fold-wise tolerant event-level precision, recall, and F1-score against Expert 1, Expert 2, and the union-based inclusive reference.

Table 6
Subject-wise strict event-level performance of the proposed framework evaluated against Expert 1, Expert 2, and Union annotations.

Subject	Expert 1			Expert 2			Union		
	P	R	F1	P	R	F1	P	R	F1
01-02-0001	0.462	0.934	0.618	0.937	0.800	0.863	0.940	0.801	0.865
01-02-0002	0.495	0.870	0.632	0.806	0.728	0.765	0.810	0.726	0.765
01-02-0003	0.222	0.959	0.360	0.860	0.879	0.869	0.870	0.881	0.876
01-02-0005	0.338	0.954	0.500	0.913	0.725	0.808	0.920	0.726	0.812
01-02-0006	0.207	0.973	0.342	0.696	0.568	0.626	0.706	0.570	0.631
01-02-0007	0.474	0.911	0.623	0.781	0.844	0.811	0.807	0.835	0.821
01-02-0009	0.543	0.940	0.688	0.930	0.778	0.847	0.941	0.776	0.851
01-02-0010	0.448	0.954	0.611	0.941	0.812	0.872	0.949	0.812	0.875
01-02-0011	0.314	0.871	0.462	0.750	0.806	0.777	0.756	0.808	0.781
01-02-0012	0.517	0.908	0.659	0.804	0.815	0.810	0.826	0.806	0.816
01-02-0013	0.363	0.914	0.520	0.718	0.860	0.783	0.730	0.855	0.787
01-02-0014	0.317	0.950	0.474	0.687	0.895	0.777	0.696	0.893	0.783
01-02-0017	0.331	0.910	0.485	0.789	0.858	0.822	0.800	0.859	0.828
01-02-0018	0.489	0.922	0.639	0.678	0.882	0.766	0.717	0.876	0.789
01-02-0019	0.261	0.966	0.411	0.744	0.811	0.776	0.749	0.814	0.780

whereas the corresponding Expert 2 and union F1-scores exceeded 0.86. Similar behavior was observed for several other subjects, indicating that Expert 1 adopted a more conservative annotation strategy.

The observed inter-subject variability is likely attributable to differences in spindle morphology, spindle density, signal quality, and annotation uncertainty. Nevertheless, the proposed framework maintained stable performance across most subjects, with union-based strict event-level F1-scores exceeding 0.75 for 13 of the 15 retained recordings. These findings demonstrate the robustness of the proposed attention-gated 1D U-Net while highlighting the continuing challenge posed by subject-specific spindle characteristics and expert annotation variability.

4.5. Agreement of model-derived spindle characteristics

To assess whether the detected events preserved basic spindle-event characteristics, we compared model-derived spindle density and duration with the corresponding expert-derived values at the subject level. Spindle density was defined as the number of spindle events per minute of analyzed N2 sleep,

$$D_s = \frac{N_s}{T_s^{\mathrm{N2}}}, \tag{38}$$

where N_s is the number of spindle events for subject s, and T_s^{N2} is the corresponding N2 duration in minutes. Event duration was computed as

$$\Delta t_j = t_j^{\mathrm{offset}} - t_j^{\mathrm{onset}}, \tag{39}$$

where t_j^{onset} and t_j^{offset} denote the onset and offset of event j, respectively. For each characteristic z, agreement between model-derived and reference-derived subject-level values was quantified using Pearson correlation coefficient r, mean absolute error (MAE), and bias. The MAE was defined as

$$\mathrm{MAE}(z) = \frac{1}{S}\sum_{s=1}^{S}\left|z_s^{\mathrm{model}} - z_s^{\mathrm{ref}}\right|, \tag{40}$$

where S is the number of evaluated subjects, z_s^{model} is the model-derived value, and z_s^{ref} is the corresponding reference-derived value.

As shown in Table 7, model-derived spindle density showed strong subject-level agreement with the union reference ($r = 0.841$) and Expert 2 ($r = 0.819$). The model density was much higher than the Expert 1 density, reflecting the substantially lower number of spindles annotated by Expert 1 and explaining the low precision but high recall observed against this reference. Duration agreement was weaker than density agreement. The model produced longer events than all reference annotations, with mean-duration biases of 0.398 s against Expert 2 and 0.396 s against the union reference. Thus, the detector captured subject-level spindle-density patterns reasonably well, particularly under the Expert 2 and union references, but showed a systematic tendency to extend event boundaries. This duration bias should be considered when using the detector for downstream analyses of spindle morphology or temporal structure.

4.6. Qualitative error analysis

To provide a more detailed qualitative assessment of the model predictions, we examined representative N2 EEG segments covering the main agreement and error modes observed in the event-level evaluation. Fig. 7 shows five examples, each including the raw EEG, sigma-filtered EEG, Expert 1 annotation, Expert 2 annotation, model probability, and final decoded prediction.

The qualitative examples support the quantitative findings by showing that the dominant error modes are not limited to complete event misses. Consensus spindles annotated by both experts were generally associated with clear sigma-band activity and elevated model probability. In contrast, single-expert events often occurred in regions with spindle-like morphology but weaker or more ambiguous temporal structure, illustrating the uncertainty introduced by manual scoring variability. Model-only detections were frequently associated with localized sigma-frequency activity that was not included in the union reference, suggesting that some false positives may correspond to plausible but unannotated spindle-like events. Missed events typically showed weaker probability responses, whereas boundary disagreements occurred when the predicted probability overlapped the annotated spindle but the decoded onset or offset differed from the manual annotation.

Overall, this qualitative analysis indicates that the model captures physiologically plausible spindle-like activity, but its errors are influenced by expert disagreement, ambiguous sigma-band transients, and event-boundary uncertainty. This observation is consistent with the higher tolerant event-level F1-score compared with strict event-level F1-score and further supports the need for expert-aware reporting in automated spindle detection.

4.7. Input channel contribution

The final model used raw EEG and sigma-band EEG as input channels. The analysis provides an approximate indication of relative channel utilization rather than a formal measure of feature importance. An exploratory inspection of the first-layer convolutional weights indicated nearly balanced normalized contributions from the raw and sigma-band inputs, with values of 0.505 and 0.495, respectively. This analysis is descriptive and should not be interpreted as formal feature attribution. It only suggests that neither input channel was strongly suppressed by the trained model.

4.8. Ablation study

Ablation experiments were conducted to quantify the contribution of the main pipeline components. All configurations were evaluated against the union-based inclusive reference. Table 8 reports the event-level results.

The largest gain was obtained by replacing candidate-based classification with dense temporal segmentation. Strict event F1 increased from 0.624 to 0.785, an absolute improvement of 16.1 percentage points. This confirms that sample-wise segmentation is better suited to spindle detection than a pipeline constrained by predefined candidate events.

Adding the sigma-band channel, probability calibration, and ensemble averaging further improved performance. The sigma channel provided frequency-specific spindle information, calibration stabilized decision thresholds across folds, and ensembling reduced sensitivity to random initialization. Together, these components increased strict event F1 to 0.799.

Table 7
Agreement between spindle characteristics derived from model predictions and expert annotations across MASS-SS2 test subjects. Agreement statistics were computed at the subject level across the five cross-validation folds.

Characteristic	Reference	Reference	Model	r	MAE	Bias
Spindle density (events/min)	Expert 1	2.727	6.389	0.872	3.662	3.662
Mean duration (s)	Expert 1	0.823	1.581	0.560	0.759	0.759
Median duration (s)	Expert 1	0.797	1.567	0.310	0.770	0.770
Spindle density (events/min)	Expert 2	6.114	6.389	0.819	0.940	0.274
Mean duration (s)	Expert 2	1.183	1.581	0.397	0.398	0.398
Median duration (s)	Expert 2	1.106	1.567	0.199	0.461	0.461
Spindle density (events/min)	Union	6.224	6.389	0.841	0.872	0.165
Mean duration (s)	Union	1.186	1.581	0.406	0.396	0.396
Median duration (s)	Union	1.107	1.567	0.222	0.460	0.460

Table 8
Ablation study of major pipeline components under the union-based inclusive reference.

Configuration	Strict P	Strict R	Strict F1	Tolerant F1
Candidate-based CNN baseline	0.662	0.607	0.624	0.791
Segmentation U-Net	0.788	0.787	0.785	0.834
Segmentation + sigma + calibration + 3-seed ensemble	0.804	0.802	0.799	0.837
Segmentation + sigma + density penalty	0.817	0.799	0.805	0.839
Segmentation + sigma + 5-seed ensemble	0.816	0.796	0.804	0.844
Segmentation + sigma + 5-seed + BCE-FPFN	0.812	0.802	0.805	0.846
Segmentation + soft labels	0.806	0.809	0.806	0.846
Final selected model	0.812	0.802	0.805	0.846

Density-aware regularization improved the precision–recall balance and increased strict F1 to 0.805. Expanding the ensemble from three to five models mainly improved tolerant event performance, suggesting more stable boundary localization. The BCE-FPFN objective achieved the best overall event-level trade-off, with strict and tolerant F1-scores of 0.805 and 0.846, respectively.

We also examined soft-label training to assess whether annotation uncertainty could be represented more explicitly. In this setting, lower-confidence regions were assigned non-binary targets rather than being treated only as hard positives or negatives. Soft labels achieved comparable event-level performance, with strict F1 of 0.806 and tolerant F1 of 0.846. However, they did not provide a consistent improvement over the final BCE-FPFN configuration and were therefore not selected as the final model.

Overall, the ablation study shows that the principal improvement comes from the dense segmentation formulation. The final framework further benefits from sigma-band input, calibration, ensemble prediction, density-aware regularization, and BCE-FPFN optimization. Compared with the candidate-based CNN baseline, the final model improved strict event F1 from 0.624 to 0.805, corresponding to an absolute gain of 18.1 percentage points and a relative improvement of 29.0%.

4.9. Novelty compared with existing deep learning spindle detectors

Recent deep learning methods, including SEED [17], SUMO [16], and SpindleU-Net [23], have advanced automated sleep spindle detection using end-to-end neural architectures. However, direct numerical comparison across these studies is limited by differences in datasets, EEG derivations, annotation references, subject selection, train–test partitions, preprocessing, and event-matching criteria. Therefore, Table 9 is intended as a methodological context table rather than a head-to-head benchmark.

The proposed framework differs from prior detectors in three main respects. First, it formulates spindle detection as dense point-wise segmentation using an attention-gated 1D U-Net with both raw EEG and sigma-band EEG inputs. This enables the model to learn broadband waveform morphology and spindle-specific oscillatory structure jointly. Second, the training and inference pipeline incorporates probability calibration, seed ensembling, and density-aware validation to improve event-level stability. Third, the evaluation explicitly reports performance against E1, E2, and the union-based inclusive reference, thereby quantifying the effect of annotation variability on reported performance.

The reported strict event-level F1-score of the proposed framework was 0.805 ± 0.031 on MASS-SS2, with a tolerant event-level F1-score of 0.846 ± 0.017. These values are within the range of previously reported deep learning spindle detectors. However, they should not be interpreted as direct superiority or inferiority relative to SEED, SUMO, or SpindleU-Net, because those methods were not re-evaluated under the same subjects, annotation references, folds, and event-matching procedure. A definitive benchmark would require running all methods under a common experimental protocol.

Accordingly, the contribution of the present study is not limited to numerical performance. The main contribution is a reproducible and annotation-aware spindle segmentation framework that combines dual EEG representations, attention-gated temporal modeling, calibration, ensemble inference, reference-wise evaluation, and zero-shot external validation.

4.10. External generalization on the DREAMS dataset

To assess external generalization, the proposed model trained exclusively on MASS-SS2 was evaluated on the DREAMS dataset without retraining, fine-tuning, or parameter adaptation. This zero-shot setting provides a rigorous assessment of cross-dataset robustness under differences in acquisition protocols, EEG montages, and annotation practices. Because the C3 derivation was not consistently available in DREAMS, C3-A1 was selected when available; otherwise, the closest central derivation (CZ-A1) was used.

Table 10 summarizes the overall performance. Under union-reference evaluation, the model achieved a strict event F1-score of 0.476 ± 0.187 and a tolerant event F1-score of 0.563 ± 0.200. Although lower than the corresponding MASS-SS2 results, the model maintained a strict recall of 0.604 ± 0.173, indicating that spindle-related patterns remained detectable despite substantial domain shift. The reduction in performance was primarily driven by decreased precision (0.426 ± 0.186), suggesting increased false-positive detections under the DREAMS annotation standard.

Notably, the sample-level ROC-AUC remained relatively high (0.873 ± 0.053), demonstrating that the learned representation preserved discriminative capability at the probability-timeline level. This observa-

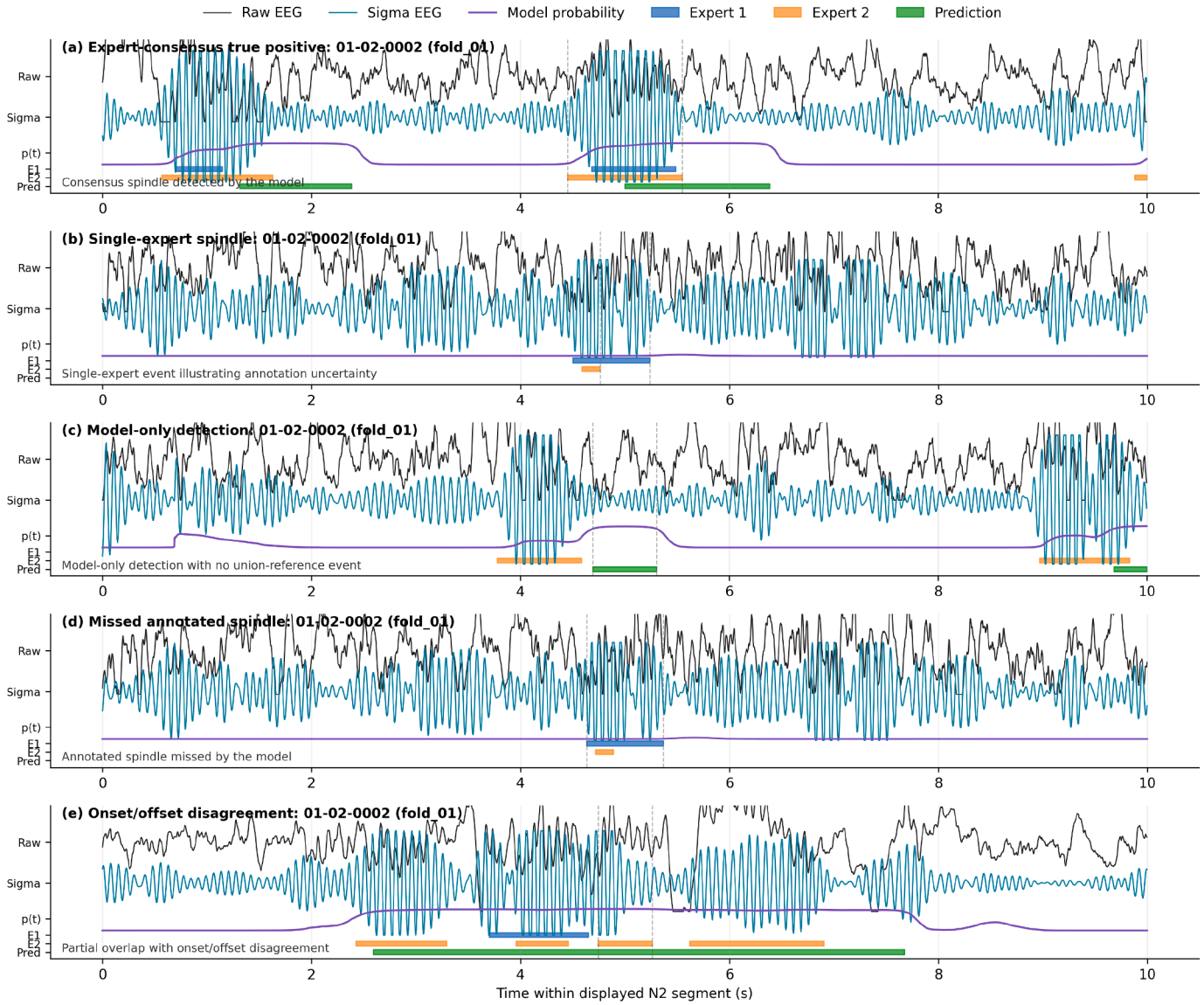

Fig. 7. Representative qualitative examples of model predictions and annotation variability. Each panel shows raw EEG, sigma-filtered EEG, model probability $p(t)$, Expert 1 annotation, Expert 2 annotation, and the final predicted events. The examples illustrate: (a) an expert-consensus spindle correctly detected by the model, (b) a single-expert spindle reflecting annotation uncertainty, (c) a model-only detection without a corresponding union-reference event at the highlighted interval, (d) a missed annotated spindle, and (e) an onset/offset disagreement with partial temporal overlap.

tion suggests that a considerable portion of the performance degradation originated from differences in event boundaries, spindle morphology, and annotation criteria rather than a complete loss of spindle localization ability.

Reference-specific results are presented in Table 11. The highest agreement was obtained against the union reference (strict F1 = 0.502), followed by Expert 2 (0.474) and Expert 1 (0.307). This trend is consistent with the MASS-SS2 analysis and further highlights the influence of annotation variability on spindle detection performance. The increase from a strict union F1-score of 0.502 to a tolerant F1-score of 0.606 indicates that many predictions were temporally close to annotated spindle events but did not satisfy the stricter overlap criterion.

Table 12 reports excerpt-wise performance. Considerable variability was observed across excerpts, with strict F1-scores ranging from 0.183 to 0.653. The strongest performance was obtained for excerpts 2, 3, 5, and 6, whereas excerpts 7 and 8 exhibited substantial over-detection, resulting in markedly reduced precision. Nevertheless, recall remained moderate across most excerpts, indicating that the learned spindle representation retained meaningful but limited transferability even under severe domain mismatch.

The performance decrease highlights the difficulty of direct cross-dataset deployment and indicates that substantial dataset-specific variability remains unresolved.

Overall, the DREAMS experiment demonstrates that the proposed attention-gated 1D U-Net learned transferable spindle-related representations and preserved meaningful detection capability under a strict zero-shot cross-dataset setting. However, the observed reduction in event-level performance highlights the challenges posed by inter-dataset variability and suggests that further improvements may require domain adaptation or target-specific calibration strategies.

5. Discussion

The proposed attention-gated dual-input 1D U-Net achieved stable subject-independent performance on MASS-SS2, with a strict event-level

Table 9
Methodological context of representative deep learning sleep spindle detectors. Reported performance values for prior studies are reproduced from the respective publications and are not directly comparable because datasets, EEG derivations, annotation references, subject partitions, preprocessing, and event-matching criteria differ across studies. NR: not reported.

Aspect	Proposed Framework	SEED [17]	SUMO [16]	SpindleU-Net [23]
Primary dataset	MASS-SS2	MASS-SS2	MODA/MASS-derived cohort	MASS-SS2
Model architecture	Attention-gated dual-input 1D U-Net	CNN–BiLSTM detector	Slim U-Net	Attention U-Net
Input representation	Raw EEG + sigma-band EEG	Raw EEG with temporal context	Single-channel EEG	Single-channel EEG
Main learning formulation	Dense point-wise spindle segmentation	Sequence-based spindle detection	Dense segmentation	Dense segmentation
Annotation reference	Expert-specific and union-based inclusive references	Expert-specific references	Study-specific manual reference	Manual annotation reference
Annotation variability analysis	E1, E2, and union evaluated separately	Partly reported through expert-specific scores	NR	NR
Uncertainty-aware training analysis	Union labels and soft-label ablation	NR	NR	NR
Calibration / ensemble strategy	Probability calibration and seed ensemble	NR	NR	NR
External validation	DREAMS zero-shot evaluation	CAP and NSRR6	NR	DREAMS
Reported strict event F1	0.805 ± 0.031	0.808–0.861	0.82	0.803 ± 0.041
Reported tolerant event F1	0.846 ± 0.017	NR	NR	NR

Table 10
Overall zero-shot external validation performance on the DREAMS dataset under union-reference evaluation.

Metric	Mean ± SD
Strict event precision	0.426 ± 0.186
Strict event recall	0.604 ± 0.173
Strict event F1	0.476 ± 0.187
Tolerant event precision	0.501 ± 0.211
Tolerant event recall	0.723 ± 0.162
Tolerant event F1	0.563 ± 0.200
Accuracy (ACC)	0.916 ± 0.018
Sensitivity (SEN)	0.502 ± 0.174
Specificity (SPE)	0.934 ± 0.016
Cohen's κ	0.266 ± 0.118
Sample F1	0.302 ± 0.125
Sample ROC-AUC	0.873 ± 0.053
Sample PR-AUC	0.206 ± 0.102

Table 11
Reference-specific pooled event-level performance on the DREAMS dataset.

Reference	Strict F1	Tolerant F1
Expert 1	0.307	0.401
Expert 2	0.474	0.547
Union	0.502	0.606

Table 12
Excerpt-wise DREAMS zero-shot performance under union-reference evaluation.

Excerpt	Strict P	Strict R	Strict F1
1	0.5448	0.5448	0.5448
2	0.5437	0.7273	0.6222
3	0.5139	0.8409	0.6379
4	0.3922	0.3175	0.3509
5	0.5547	0.6893	0.6147
6	0.6183	0.6923	0.6532
7	0.1200	0.6667	0.2034
8	0.1232	0.3542	0.1828

F1-score of 0.805 ± 0.031 and a tolerant F1-score of 0.846 ± 0.017 under the union-based inclusive reference. The ablation analysis showed that the principal improvement resulted from reformulating spindle detection as dense temporal segmentation, increasing strict F1 from 0.624 to 0.805 relative to the candidate-based baseline. This supports sample-wise probability estimation as an effective strategy for localizing short and temporally variable spindle events. The combined use of raw and sigma-band EEG further provided complementary morphological and frequency-selective information, although the first-layer weight analysis should be interpreted only as descriptive evidence of channel utilization rather than formal feature attribution.

Annotation variability had a substantial effect on measured performance. Strict event-level F1 was 0.563 against Expert 1, 0.800 against Expert 2, and 0.805 against the union reference. The lower agreement with Expert 1 was driven mainly by reduced precision, consistent with the markedly lower number of events annotated by this scorer. Thus, many detections accepted by Expert 2 or by the inclusive union reference were counted as false positives when Expert 1 was used alone. These findings show that detector performance is conditional on the annotation standard and support reference-specific reporting when inter-scorer disagreement is substantial.

The union reference should therefore be interpreted as an inclusive supervisory target rather than expert consensus. It avoids treating all single-rater events as negatives, but it also assigns equal positive status to events supported by one or both experts and may consequently reflect the more inclusive scorer. The present soft-label experiment provided an initial assessment of uncertainty-aware supervision, but did not produce a consistent improvement over the selected BCE-FPFN configuration. More rigorous comparison of union, intersection, expert-specific, and probabilistic targets is needed to determine how annotation uncertainty should be represented during training.

External validation on DREAMS exposed a substantial generalization gap. Strict event-level F1 decreased from 0.805 on MASS-SS2 to 0.476 under zero-shot evaluation. This decline is consistent with combined shifts in cohort characteristics, EEG derivation, acquisition settings, and annotation practice. The retained sample-level discrimination indicates that spindle-related information was not completely lost, but the reduced event-level agreement shows that decision thresholds, waveform characteristics, and event boundaries remain dataset dependent. The DREAMS experiment should therefore be interpreted as evidence of partial zero-shot transfer rather than reference-invariant or clinically robust generalization.

Finally, agreement with manual annotations should not be equated with physiological validity. High event-level F1 does not guarantee unbiased estimation of spindle density, duration, amplitude, or frequency. Such quantities are often used in studies of memory, aging, medica-

tion effects, and neurological disease. The present framework should therefore be regarded primarily as a technical approach for automated N2 spindle segmentation under the evaluated conditions. Its suitability for downstream biomarker analysis requires direct validation of model-derived spindle characteristics and their relationships with clinically or cognitively meaningful outcomes.

5.1. Limitations and future work

Several limitations define the scope of the present findings. First, all analyses were restricted to N2 sleep; the conclusions should therefore not be generalized to other NREM stages without additional evaluation. Second, the training cohort was predominantly healthy. Spindle morphology may vary with age, medication, neurological diagnosis, sleep architecture, and background EEG characteristics. Abnormal clinical recordings may additionally contain epileptiform discharges, sharp transients, slowing, and artifacts. These patterns may be particularly challenging for the sigma-band input because filtering sharp activity can generate oscillatory components within the spindle-frequency range.

Third, the use of a single central EEG derivation limits spatial characterization of spindle activity. Regional differences between frontal slow spindles and centroparietal fast spindles cannot be fully represented by a single channel. Multi-channel and spatially informed models should therefore be investigated. EEG referencing is another important source of domain shift, because re-referencing can alter signal amplitude, polarity, spectral contrast, and signal-to-noise ratio. Future work should examine re-reference augmentation, multi-reference training, and reference-robust representations.

Future studies should also incorporate explicit uncertainty-aware supervision using intersection, expert-specific, probabilistic, or multi-rater labels. Validation should extend to heterogeneous clinical cohorts and artifact-rich EEG, with direct comparison of predicted and manually derived spindle density, duration, amplitude, and frequency. Cross-study performance comparisons should remain cautious unless competing methods are evaluated using identical subjects, folds, preprocessing, annotation references, and event-matching criteria. A shared benchmark under a common evaluation protocol would provide a more rigorous basis for assessing methodological improvements.

6. Conclusion

This study presented an attention-gated dual-input 1D U-Net framework for automated N2 sleep spindle segmentation from single-channel EEG. By reformulating spindle detection as dense temporal segmentation and combining raw EEG with sigma-band input, the proposed approach improved event localization compared with the candidate-based CNN baseline. Ensemble prediction, probability calibration, density-aware validation, and BCE-FPFN optimization further supported stable event-level performance.

The results also show that annotation choice has a major effect on reported performance. The model achieved strong agreement with the union-based inclusive reference and Expert 2, but lower precision against the more conservative Expert 1. The union reference should therefore be interpreted as an inclusive composite annotation, not as expert consensus. Reporting E1-, E2-, and union-based results provides a more transparent assessment of detector behavior under inter-scorer variability.

External evaluation on DREAMS showed partial zero-shot discrimination under cross-dataset and cross-reference domain shift. However, these findings do not establish reference-invariant or clinically reliable spindle detection. The experiments were restricted to N2 sleep and were based mainly on healthy training recordings. Broader validation across abnormal clinical EEG, neurological cohorts, medication profiles, recording systems, EEG references, and additional sleep stages is required before clinical deployment.

CRediT authorship contribution statement

Waqas Ahmed: Conceptualization, methodology, software, formal analysis, investigation, data curation, validation, visualization, and writing—original draft. **Keijo Haataja**: Supervision, methodology, project administration, funding acquisition, resources, writing—review and editing. **Pekka Toivanen**: Supervision, methodology, project administration, resources, writing—review and editing.

Consent for publication

Not applicable. The manuscript does not contain any individual participant information, identifiable images, or personal data requiring consent for publication.

Ethics approval and consent to participate

This study did not involve the recruitment of human participants, collection of new human data, or any intervention involving human subjects. All experiments were conducted using the publicly available Montreal Archive of Sleep Studies (MASS-SS2) and DREAMS sleep spindle datasets, which consist of fully de-identified polysomnographic recordings.

Human and animal rights

No experiments involving humans or animals were conducted by the authors. The study exclusively analyzed previously collected and anonymized public datasets.

Funding

This research work was funded by the EU KDT H2TRAIN Project (Grant Agreement Number: 101140052), supported by the Chips Joint Undertaking and its members, including top-up funding by the National Funding Agency for Finland, the Innovation Funding Agency Business Finland.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge the developers and contributors of the MASS and DREAMS databases for making these valuable datasets publicly available, enabling reproducible research in automated sleep spindle detection.

Data availability

The datasets analyzed during this study are publicly available:

- Montreal Archive of Sleep Studies (MASS-SS2)
- DREAMS Sleep Spindle Database

References

[1] W. Ahmed, P. Toivanen, K. Haataja, Towards sleep spindle detection: a comparative survey of state-of-the-art, challenges and future research, IEEE Access 13 (2025) 182821–182845. https://doi.org/10.1109/ACCESS.2025.3622316.
[2] F. Wang, L. Li, Y. Wan, Z. Li, L. Luo, B. Hu, J. Pan, Z. Wen, H. Huang, An efficient sleep spindle detection algorithm based on MP and LSBoost, Comput. Mater. Contin. 76(2) (2023) 2301–2316.
[3] U. Hassan, G.B. Feld, T.O. Bergmann, Automated real-time EEG sleep spindle detection for brain-state-dependent brain stimulation, J. Sleep Res. 31(6) (2022) e13733.
[4] K.A. Paller, J.D. Creery, E. Schechtman, Memory and sleep: how sleep cognition can change the waking mind for the better, Annu. Rev. Psychol. 72(1) (2021) 123–150.

[5] W. Duan, Z. Xu, D. Chen, J. Wang, J. Liu, Z. Tan, X. Xiao, P. Lv, M. Wang, K.A. Paller, et al., Electrophysiological signatures underlying variability in human memory consolidation, Nat. Commun. 16(1) (2025) 2472.
[6] A. Castelnovo, A. D'Agostino, A. Mayeli, L. Albantakis, G. Tononi, F. Ferrarelli, Sleep spindle abnormalities as neurophysiological biomarkers of schizophrenia spectrum disorders: from cellular mechanisms and neural circuits to clinical implications, Biol. Psychiatry 98(11) (2025) 809–818.
[7] J. Kim, M. Park, Systems memory consolidation during sleep: oscillations, neuromodulators, and synaptic remodeling, BMB Rep. 58(10) (2025) 425.
[8] M.E. Dimitriades, A. Markovic, S.R. Gefferie, A. Buckley, D.I. Driver, J.L. Rapoport, M. Nosadini, K. Rostasy, S. Sartori, A. Suppiej, et al., Sleep spindles across youth affected by schizophrenia or anti-N-methyl-D-aspartate-receptor encephalitis, Front. Psychiatry 14 (2023) 1055459.
[9] B. Baran, E.E. Lee, Age-related changes in sleep and its implications for cognitive decline in aging persons with schizophrenia: a critical review, Schizophr. Bull. 51(2) (2025) 513–521.
[10] F. Ferrarelli, Sleep spindles as neurophysiological biomarkers of schizophrenia, Eur. J. Neurosci. 59(8) (2024) 1907–1917.
[11] J.J. Eggermont, Brain Oscillations, Synchrony and Plasticity: Basic Principles and Application to Auditory-Related Disorders, Academic Press, 2021.
[12] M. Abdul Ghani, Development and Validation of a Machine Learning Pipeline for Automating Canine Sleep Staging Under Sedated and Normal Sleep, Ph.D. thesis, University of Guelph, 2025.
[13] S. Zhou, G. Song, H. Sun, D. Zhang, Y. Leng, M.B. Westover, S. Hong, Continuous sleep depth index annotation with deep learning yields novel digital biomarkers for sleep health, NPJ Digit. Med. 8(1) (2025) 203.
[14] E. Ramirez, P.A. Estevez, M.D. Adams, C.A. Perez, M. Garrido Gonzalez, P. Peirano, USSD: unsupervised sleep spindle detector, IEEE Access 13 (2025) 18644–18659. https://doi.org/10.1109/ACCESS.2025.3532536
[15] I.A. Zapata, P. Wen, E. Jones, S. Fjaagesund, Y. Li, Automatic sleep spindles identification and classification with multitapers and convolution, Sleep 47(1) (2024) zsad159.
[16] L. Kaulen, J.T.C. Schwabedal, J. Schneider, P. Ritter, S. Bialonski, Advanced sleep spindle identification with neural networks, Sci. Rep. 12(1) (2022) 7686.
[17] N.I. Tapia-Rivas, P.A. Estévez, J.A. Cortes-Briones, A robust deep learning detector for sleep spindles and K-complexes: towards population norms, Sci. Rep. 14(1) (2024) 263.
[18] D. Jiang, Y. Ma, Y. Wang, A robust two-stage sleep spindle detection approach using single-channel EEG, J. Neural Eng. 18(2) (2021) 026026.
[19] J. Pan, Z. Yang, Q. Shen, M. Li, C. Jiang, Y. Li, Y. Li, Deep learning-augmented sleep spindle detection for acute disorders of consciousness: integrating cnn and decision tree validation, IEEE Trans. Biomed. Eng. 72(10) (2025) 3120–3132. https://doi.org/10.1109/TBME.2025.3562067.
[20] J. Rychlík, R. Mouček, Automatic detection of sleep spindles by neural networks algorithms, Acta Polytech. CTU Proc. 51 (2024) 75–80.
[21] L. Wei, S. Ventura, M.A. Ryan, S. Mathieson, G.B. Boylan, M. Lowery, C. Mooney, Deep-spindle: an automated sleep spindle detection system for analysis of infant sleep spindles, Comput. Biol. Med. 150 (2022) 106096.
[22] Z. Yang, J. Pan, Automatic detection of sleep spindles and its application in patients with acute disorders of consciousness, in: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2023, pp. 733–738.
[23] J. You, D. Jiang, Y. Ma, Y. Wang, SpindleU-Net: an adaptive U-Net framework for sleep spindle detection in single-channel EEG, IEEE Trans. Neural Syst. Rehabil. Eng. 29 (2021) 1614–1623.
[24] J. Tian, Y. Wei, Sleep spindle auto detection based on U-Net with label optimization module, in: Proceedings of the 2025 2nd International Conference on Generative Artificial Intelligence and Information Security, 2025, pp. 382–391.
[25] P. Chen, D. Chen, L. Zhang, Y. Tang, X. Li, Automated sleep spindle detection with mixed EEG features, Biomed. Signal Process. Control 70 (2021) 103026.
[26] T. Hu, Z. Wang, J. Liu, T. Yang, W. Chen, A biLSTM-based framework for sleep spindle and K-complex detection with enhanced segment-level learning, IEEE Trans. Biomed. Eng. (2026) 1–10. https://doi.org/10.1109/TBME.2026.3679596.
[27] V.S. Babu, A. Ramakrishna, D.S. Rao, Automatic sleep spindle detection using SMOTE and composite features with SWT and adaboost, Int. J. Intell. Eng. Syst. 18(1) (2025) 1173–1186.
[28] L. Kaulen, J.T.C. Schwabedal, J. Schneider, P. Ritter, S. Bialonski, SUMO: advanced sleep spindle identification with neural networks, arXiv e-prints (2022) arXiv–2202. https://doi.org/10.1038/s41598-022-11210-y.
[29] E. Gurdiel, F. Vaquerizo-Villar, J. Gomez-Pilar, G.C. Gutiérrez-Tobal, F. Del Campo, R. Hornero, Beyond the ground truth, XGBoost model applied to sleep spindle event detection, IEEE J. Biomed. Health Inform. 29(7) (2025) 4873–4883. https://doi.org/10.1109/JBHI.2025.3544966.
[30] S. Devuyst, The DREAMS databases and assessment algorithm, 2005, https://doi.org/10.5281/zenodo.2650142
[31] C. O'reilly, N. Gosselin, J. Carrier, T. Nielsen, Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research, J. Sleep Res. 23(6) (2014) 628–635.