Beyond the ground truth, XGBoost model applied to sleep spindle event detection

Enrique Gurdiel, Fernando Vaquerizo-Villar, Javier Gomez-Pilar, *Member, IEEE*, Gonzalo C. Gutiérrez-Tobal, *Member, IEEE*, Félix del Campo, and Roberto Hornero, *Senior Member, IEEE*.

***Abstract*— Sleep spindles are microevents of the electroencephalogram (EEG) during sleep whose functional interpretation is not fully clear. To streamline the identification process and make it more replicable, multiple automatic detectors have been proposed in the literature. Among these methods, algorithms based on deep learning usually demonstrate superior accuracy in performance assessment up to now. However, using these methods, the rationale behind the model decision-making process is hard to understand. In this study, we propose a novel machine-learning detection framework (SpinCo) based on an exhaustive sliding window feature extraction and the application of XGBoost algorithm, achieving performance close to state-of-the-art deep-learning techniques while depending on a fixed set of easily interpretable features. Additionally, we have developed a novel by-event metric for evaluation that ensures symmetricity and allows a probabilistic interpretation of the results. Through the utilization of this metric, we have enhanced the interpretability of our evaluations and enabled a direct assessment of inter-expert agreement in the manual annotation of spindle events. Finally, we propose a new type of performance assessment test based on estimations of the automatic method's ability to generalize to unseen experts and its comparison with inter-expert agreement measurements. Hence, Spinco is a robust automatic spindle detection technique that can be used for labeling raw EEG signals and shed light on the metrics used for evaluation in this problem.**

***Index Terms*— Electroencephalography, machine learning, polysomnography, signal processing, sleep spindle detection, XGBoost.**

I. Introduction

SPINDLES are bursts of electroencephalographic (EEG) activity caused by thalamocortical interactions during sleep. In humans, they are distinctive of sleep stage N2, although also produced during N3 [1]. These sleep microevents have a duration of at least 0.5 seconds and are localized in frequency, typically between 11 and 16 Hz [2]. Nevertheless, that range can vary depending on the characteristics of the population under study, such as age [3]. Relationships have been described between higher density of spindles (number of spindle events per hour of N2 and N3 sleep) and lesser propensity to develop insomnia during stressful situations [4], as well as with intellectual ability measurements [5], [6], and overnight memory consolidation [7]. Furthermore, statistically significant alterations of the spindle properties have been described in subjects with sleep apnea [8], Alzheimer's disease [9], or schizophrenia [10]. Together, these findings support the functional interpretation of spindles as mechanisms of sleep maintenance, biological markers of intellectual development, helpers of memory consolidation, or cerebral plasticity facilitators [11].

Identification of this type of microevents is traditionally done by visual inspection and annotation conducted by experts [12]. This is a time-consuming task of low replicability [13] that has led to the appearance of multiple automatic detection methods in the literature. Three principal categories of automatic spindle detectors can be distinguished. The first category is composed of signal processing methods, which depend on the definition of one or multiple features that can be calculated in each sample of the signal at the original sample rate. These features are then thresholded individually to obtain candidates for spindle detection that are then processed. In this category, we can find methods depending on a single feature as the envelope of the sigma band [10], [12], [14] or more advanced techniques depending on multiple spectral and nonlinear features [15]. Another variation of these methods is based on the decomposition of the signal into an oscillatory and transient component [16], [17]. The main advantage of this category is that it can be easily implemented as a sample-based model, thus not losing temporal accuracy in the spindles' onset and offset. Still, these methods underperform compared with more advanced techniques. The second category is machine

F. Vaquerizo-Villar was supported by a Sara Borrell grant (CD23/00031) from the ISCIII cofounded by the Fondo Social Europeo Plus (FSE+). GC. Gutiérrez-Tobal was supported by a postdoctoral grant from the University of Valladolid. This work is part of the projects PID2020-115468RB-I00, PID2023-148895OB-I00, and CPP2022-009735, funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR. This research was also co-funded by the European Union through the Interreg VI-A Spain-Portugal Program (POCTEP) 2021-2027 (0043_NET4SLEEP_2_E), and by Consorcio Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN) under grant CB19/01/00012 through Instituto de Salud Carlos III (ISCIII), in part by European Regional Development, and under the project Tattoo4Sleep from 2022 CIBER-BBN Early Stage Plus call. *(Enrique Gurdiel and Fernando Vaquerizo-Villar contributed equally to this work). (Corresponding authors: Fernando Vaquerizo-Villar and Javier Gómez Pilar.)*

Enrique Gurdiel, Javier Gomez-Pilar, Gonzalo C. Gutiérrez-Tobal, Félix del Campo, and Roberto Hornero are with the Biomedical Engineering Group, Universidad de Valladolid, 47011 Valladolid, Spain, and also with the CIBER-BBN, ISCIII, 28029 Madrid, Spain (e-mail: enrique.gurdiel@gib.tel.uva.es; javier.gomez.pilar@uva.es; gonzalocesar.gutierrez@uva.es; fsas@telefonica.net; roberto.hornero@uva.es).

Fernando Vaquerizo-Villar is with the Biomedical Engineering Group, Universidad de Valladolid, 47011 Valladolid, Spain, also with the CIBER-BBN, ISCIII, 28029 Madrid, Spain, and also with the Anaesthesiology Department, Hospital Clínico Universitario de Valladolid, 47003 Valladolid, Spain (e-mail: fernando.vaquerizo@uva.es).

learning (ML) systems. In this category, the signal is usually separated into epochs of a given window size, subsequently extracting particular features in each epoch. Then, a ML algorithm endeavors to predict for each epoch whether it contains a spindle or not [18]–[21]. This kind of approach is generally more accurate than the previous category, but they have the pitfall that some heuristics must be performed to define onsets and offsets of spindles. Finally, the third category is composed of deep learning (DL) approaches, typically inspired by image object detection frameworks [22]–[26]. While these techniques have obtained outstanding performance in benchmark tests [22], [25], [26], they sacrifice the ability to easily interpret the decision-making process when detecting events. Our proposal aims to harness the advantages of all three categories, including high temporal resolution and cutting-edge predictive capabilities, while still retaining the ability to maintain interpretability. Thus, we aim to achieve a well-performing system that depends on an explicit set of interpretable features. The interested reader can further explore other spindle detection algorithms in [22], [26]–[28].

Many efforts have been made in the problem of spindle event detection to improve the evaluation metrics available. One possibility is to use a by-sample approach in which the confusion matrix is calculated based on the individual samples of a signal to compute the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [29]. For a complete explanation of the different by-sample metrics available and their use in the literature, see [19]. This approach has the drawback of giving more importance to long than short spindles as the number of samples varies [27]. The introduction of event-based metrics enables a one-to-one comparison of spindles [27], providing a more interpretable value and establishing itself as the standard evaluation method in recent years [15], [17], [19], [22], [26]. Consequently, we assess our methodology using by-event metrics.

Another major concern within the research community involved in automatic spindle detection arises from the observed low inter-expert agreement. In response, some researchers have opted to augment the number of experts and subjects involved in experiments, facilitating a crowdsourced approach to define the ground truth (or gold standard) [15], [27], [30], [31] Nevertheless, those approaches are based on labelling small sections of EEG signals instead of entire nights, and each expert labels only a reduced part of those sections at random, what makes the type of analysis we are proposing unfeasible. In this study, we chose a novel approach to refine the by-event metric that would make it symmetric for the interchange of ground truth and tested annotations, interpretable as a similarity metric, and robust to previously ill-defined cases. This would give us the chance to fairly explore inter-expert agreement even with just two experts available as we can isolate the two main sources of variability for spindle scoring, the expert annotating and the subject recorded. Advances in the event detection metrics could be of interest in other medical detection tasks [32]. We encounter that different estimations or inter-expert agreement are present in the literature. In [33], authors estimated an 86% interscorer reliability, while in the studies based on crowdsourced definitions of a ground truth [27], [31], this reliability is indirectly estimated by the average expert F1 score compared to the crowdsourced ground truth, which is approximately 0.7. In direct estimations in the MASS SS2 database [34], the F1 score only reaches 0.54. However, this may be an underestimation due to differences in the scoring followed by each of the two experts involved in the study [12].

The objective of this research is twofold. First, to develop a ML EEG microevent detection framework, called SpinCo, which performs at levels comparable to state-of-the-art methods in the problem of sleep spindle event detection, while remaining easily interpretable. Second, to further develop the by-event metric used in the literature to make it symmetric and robust to previously ill-defined situations. To achieve these objectives, a procedure is employed that utilizes three distinct databases, thereby validating the robustness of the model, along with the XGBoost algorithm [35], a gradient-boosted trees method that has been previously used in diverse medical applications [36], [37].

In summary, we propose in our research a robust ML framework, SpinCo, validated in multiple databases. We have introduced novelties in the feature selection process and in the metric used for the evaluation of spindle event detection.

II. Materials and methods

A. Databases

In this study, three different sources of EEG recordings with spindle data labeled by experts are used: two public databases of EEG signals with spindle annotations by experts of widespread use (DREAMS [38] and MASS [34]), and a private database from Río Hortega University Hospital (COGNITION). DREAMS spindle database includes 8 EEG annotated excerpts of 30 minutes, 2 from channel C3 and 6 from CZ, of adult subjects with several pathologies (age 45.9 ± 7.6). Among them, 6 recordings were annotated by two experts independently while the remaining 2 were annotated only by one of them. MASS second subsection (SS2) includes 19 complete overnight signals, all from healthy adult subjects (age 23.6 ± 3.7). Among them, 15 signals were annotated by two experts while the remaining 4 were annotated only by one, all in channel C3. COGNITION database includes 7 complete overnight signals, all from pediatric subjects with sleep-related pathologies (age 7.9 ± 1.0, BMI 18.1 ± 2.6). All signals were annotated in the C3 channel by a single expert. Additional characteristics regarding spindles in the different databases are shown in Table I. This work has been carried out according to the Declaration of Helsinki. The research protocol has been validated by the Ethics Committee of the Río Hortega University Hospital of Valladolid (Ref.: 22-PI159), and the legal caretakers of the children signed a written informed consent when collecting data for our private database. From the values of Table I, it is noteworthy that the database spindle density of COGNITION is significantly lower than that of DREAMS or MASS, which can be explained by the lower spindle density of infants compared to adults [3].

TABLE I
SPINDLE STATISTICS BY DATABASE.

Database	Number of subjects	Number of experts	Recording time[a] (hours)	Spindle density (epm)	Total spindle annotations
DREAMS	8	2	0.5 ± 0.0	1.48[b] and 2.48[c] (E1[d]), 2.27[b] and 4.04[c] (E2[e])	764
MASS	15	2	8.2 ± 0.9	1.07[b] and 2.77[c] (E1[d]), 2.35[b] and 6.00[c] (E2[e])	33458
COGNITION	7	1	11 ± 0.0	0.53[b] and 1.13[c]	2785

epm: events per minute. [a]Mean ± standard deviation among subjects. [b]All the recording. [c]N2 sleep. [d]E1 = expert one. [e]E2 = expert two

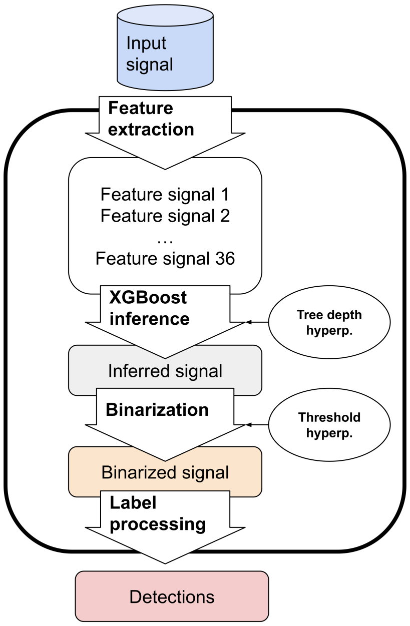

Fig. 1. **Schematic Representation of the SpinCo Detection Framework.** In the first stage the signal features are extracted, then the XGBoost model infers a signal of spindle likelihood that is later binarized with a single threshold. Finally, the binarized signal is processed to obtain the output annotations, or detections, of our automatic system.

B. Spindle detection framework: SpinCo

Fig. 1 shows the overall SpinCo pipeline for the detection of EEG microevents and its application to spindle detection. This approach is based on an exhaustive sliding window feature extraction, using a step of one sample and appropriate padding so each feature is of the same dimension as the original signal. In this way, we train a model for spindle detection that outputs for each sample a value of likelihood to be part of a spindle. A threshold is applied to that output and is then processed to obtain the detections, as explained in the following subsections. Importantly, each stage is executed concurrently for the whole signal, eliminating the need for defining specific epochs. Additionally, no artifact rejection is applied in the process, creating a system that can learn to reject the artifacts automatically. The details of the implementation of all the stages of SpinCo can be consulted in our public and free available repository [39].

1) Feature extraction: When a ML system depends on an explicit feature extraction, there is a risk of introducing bias with the initial proposal of features, as these may depend on the frequency bands or window sizes considered [22]. To avoid this bias, SpinCo involves complementary linear and non-linear analyses. Specifically, we work with generic features of spectral power (Hjort activity), mean frequency (Hjort mobility), complexity (Hjort complexity), and fractal dimension (Petrosian fractal dimension) extracted in 8 spectral bands that reflect the domain knowledge [40] (delta1 from 0.1 to 2 Hz, delta2 from 2 to 4 Hz, theta from 4 to 8 Hz, alpha from 8 to 13 Hz, sigma from 11 to 16 Hz, beta1 from 13 to 19 Hz, beta2 from 19 to 30 Hz and broadband from 0 to 40 Hz), using 4 different window sizes (0.5, 1, 1.5, and 2 s) for feature extraction. Additionally, in accordance with the results of the literature, the well-known spindle-specific feature, sigma index, is included [18], [20], [41], [42]. A sliding windows approach has been previously used in [15], [43], [44], where a step of 0.1s, 0.125s, or 0.5s was applied respectively. This approach was also used in [18], [21] with a step of 1 sample and fixed window size. To the best of our knowledge, no previous method has extracted features at windows of different sizes simultaneously. Allowing the model to learn from different window sizes, we leave open the possibility of selecting particular scales for a given feature, reducing the bias in the feature selection process.

Before the feature extraction stage, the signals are preprocessed in three steps: (i) resampling to 200Hz to ensure a fixed and sufficiently high sample rate for spindle identification; (ii) applying a 4th order Butterworth low pass filter at 40 Hz to avoid the power line frequency; and (iii) applying robust z-score normalization to the signal. The z-score is calculated by removing the median and scaling with the interquartile range computed across the entire recording, which makes it more robust to outliers than the standard z-score normalization. The generic features calculated for all the spectral bands considered were the Hjort parameters (activity, mobility, and complexity), previously used in [19], [20], and the Petrosian fractal dimension, proposed in [45] and used in [20]. Considering the signal as $y(t)$, where t is time, we can describe the features as follows:

Hjort activity:

$$activity = variance\left(y\left(t\right)\right). \tag{1}$$

Hjort mobility:

$$mobility = \sqrt{\frac{variance\left(\frac{dy(t)}{dt}\right)}{variance\left(y\left(t\right)\right)}}. \tag{2}$$

Hjort complexity:

$$complexity = \frac{mobility\left(\frac{dy(t)}{dt}\right)}{mobility\left(y\left(t\right)\right)}. \tag{3}$$

Petrosian fractal dimension:

$$petrosian = \frac{log_{10}\left(w\right)}{log_{10}(w) + log_{10}(w/(w+0.4\partial)}, \tag{4}$$

where w is the number of samples in the window considered, and ∂ is the number of sign changes of the first temporal derivative of the signal in that window.

The specific spindle feature is defined as follows:

Sigma index:

$$sigma\ index = \frac{mean\left(\left|B3\left(t\right)\right|\right)}{mean\left(\left|B1\left(t\right)\right|\right) + mean\left(\left|B2\left(t\right)\right|\right)}, \tag{5}$$

where $B1$, $B2$, and $B3$ are the signal filtered in the frequency intervals 4 to 10 Hz, 20 to 40 Hz and 12.5 to 15 Hz, respectively [20]. In total, we extract the 4 general features at 8 different spectral bands and 4 different window sizes, as well as the spindle specific feature at 4 different window sizes, a total of 132 features. Hence, apart from the spindle specific feature, our feature extraction process contains general information of the EEG that can be related not only to the presence of spindles but also to characterize its absence, so the feature selection method is executed in a low biased set of features and its result can be interpreted afterwards.

2) Feature selection: In order to reduce computational burden during training and inference on new data, as well as to minimize overfitting, a feature selection technique is applied. We use an embedded technique in XGBoost applied to the whole DREAMS database, together with a bootstrapping procedure with $N = 1000$ repetitions to obtain an optimum subset of features that are reliably selected.

For each bootstrap round, XGBoost was first trained using 30 boosting iterations in which the union of the labels of both experts is used for ground truth, a typical criterion when using DREAMS in the literature [19], [24], [46]. Then, we ordered the feature by the ratio of tree splits and kept those features whose cumulative ratio of tree splits is higher than 50%, as represented in Fig. 2. This process is repeated N times, so we obtain, for each feature, the number of times it was selected (N_{sel}). Those variables that were selected for at least 25% of the runs (250) formed the optimum subset, as we experimentally determined that they maximized our results. The selected features can be consulted in the results section.

3) Classification: The algorithm used in this research is XGBoost [35], an efficient implementation in Python of gradient boosting chosen by its robustness to feature redundancy and imbalance problems. During the training phase, we train a high enough number of boosting iterations, 60 in our case, to later optimize the number of boosting iterations used for inference in validation, as this is the main parameter controlling overfitting. We train with a binary logistic objective function. We choose a learning rate of 0.4 heuristically, with an optimal trade-off between training speed and the convergence value of the metric. We used a subsampling of 0.6 that speeds up training while minimizing overfitting, and ensuring it is above 0.5 as recommended [35]. The rest of the hyperparameters were left to the default parameters to simplify the validation, as heuristically no further improvement has been observed by altering them.

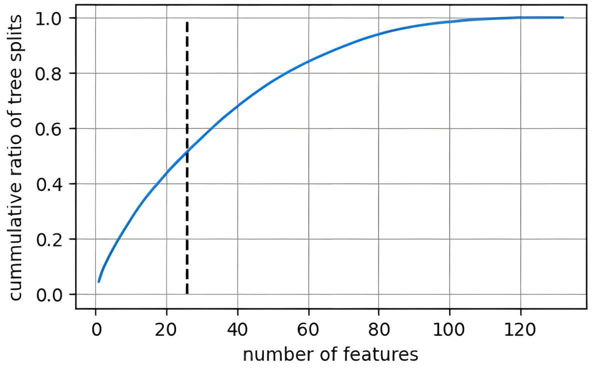

Fig. 2. **Feature selection overview.** Number of features selected against the cumulative ratio of tree splits in the XGBoost embedded feature selection method.

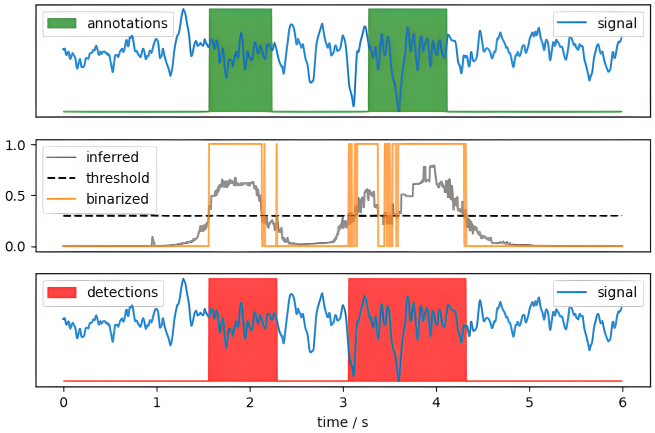

Fig. 3. **Representation of the binarization and label processing stages and comparison with original annotations.** In blue, a 6 second excerpt of signal in arbitrary units. In green, expert annotations. In grey, the inferred signal after XGBoost model is applied. In black dashed, a binarization threshold of 0.3. In orange, the binary signal obtained. In red, the detections after the label processing stage.

4) Binarization: The binarization consists of the transformation of the model output, a real-valued signal between 0 and 1, in a binary signal that can be interpreted as a logical value of whether the model has detected a spindle in a specific sample or not. This step is performed by applying a single threshold value between 0 and 1 that is optimized in the validation stage, as it is the main hyperparameter controlling the trade-off between precision and recall. A representative example of this process can be found in Fig. 3.

5) Label processing: In the latter step of our method, detections that are closer than 0.1 seconds are joined together [15], [47], and detections shorter than 0.3 seconds are discarded [27], [48]. In the particular case of the second expert of MASS SS2, we removed annotations larger than 5 seconds, as this expert had annotated 7 unusually long spindles. In order to apply these operations efficiently, we used image processing operations, closing, and object detection, of the multidimensional image processing SciPy python package.

6) Validation procedure: In our experiments, leave-one-subject-out cross-validation (LOOCV) in the MASS and COGNITION databases is used. In MASS, we used three subjects at random for each testing subject for validation purposes, while in the COGNITION database, we used only two due to the lesser number of subjects. In MASS, we trained and validated for expert 1 (E1) and expert 2 (E2) separately, considering only the 15 subjects that are labeled by both experts to use the same training, validation, and test sets and allow for direct comparison of the results. For each fold, SpinCo hyperparameters are set by grid search on the validation sets as in [26], optimizing the mean over the validation subjects of the metric $F1^*$, which is explained in the following section, using intersection over union (IoU) thresholds of 0.2 [26] and 0.3 [22].

C. Evaluation methodology

We here develop the modifications in the way the by-event metric is calculated to make it symmetric (hence applicable to the general comparison of two observations without the need of assuming one is the ground truth), robust to cases in which the group of intersecting spindles is larger than 2, and interpretable in terms of probability. The calculation of a by-event metric is based on an auxiliary metric, IoU, which quantifies the degree of overlap between two given annotations as a number between 0 (no intersection) and 1 (total overlap) [27]. The IoU metric is calculated, for two given detections, as the ratio of the number of samples pertaining to both of them (intersection) divided by the number of samples pertaining to at least one of them (union). A IoU threshold value, x, is set in this auxiliary metric to define the TP value. In order to calculate it, a helpful operation is the definition of a IoU matrix that comprises all the possible pairs of annotations of the two observations that are being compared:

$$IoUmatrix = \begin{pmatrix} IoU_{11} & ... & IoU_{1M} \\ ... & ... & ... \\ IoU_{N1} & ... & IoU_{NM} \end{pmatrix}, \tag{6}$$

where we are comparing an observation of N annotations (row index) with another of M annotations (column index). We consider, without loss of generality, that the row index represents the ground truth annotations, and we operate taking the maximum by row, as in (7). Thus, we can calculate the number of TP as the number of rows in which the maximum value of IoU exceeds the threshold x. In (8), we name this value as $TP1$.

$$recallValues = \begin{pmatrix} max(IoU_{11}, & ..., & IoU_{1M}) \\ & ... & \\ max(IoU_{N1}, & ..., & IoU_{NM}) \end{pmatrix}. \tag{7}$$

$$TP1 = sum(recallValues > x). \tag{8}$$

With this approach, the recall is defined as the probability of a ground truth annotation to be TP straightforwardly, as per (9).

$$recall = \frac{TP1}{N}. \tag{9}$$

In previous proposals of an event-metric [22], [26], [27], the number of TP are calculated in this way, while the rest of the values of the confusion matrix needed to define $F1$ are obtained with arithmetical operations as $FN = N - TP$ and $FP = M - TP$. Nevertheless, we can operate similarly by column, as in (10), and define another value of TP, named $TP2$ in (11).

$$precisionValues = \left(max\begin{pmatrix} IoU_{11} \\ ... \\ IoU_{N1} \end{pmatrix}, \quad ..., \quad max\begin{pmatrix} IoU_{1M} \\ ... \\ IoU_{NM} \end{pmatrix} \right). \tag{10}$$

$$TP2 = sum(precisionValues > x). \tag{11}$$

Then, the precision can be defined as the probability of an annotation generated by the automatic method to be TP, as per (12).

$$precision = \frac{TP2}{M}. \tag{12}$$

If we consider $TP1$ the number of true positives of the N ground truth annotations and $TP2$ the number of true positives of the M annotations generated by an automatic method, we can define the probability of an annotation to be true positive, named as $F1^*$ in (13).

$$F1^* = \frac{TP1 + TP2}{N + M} = \frac{2 \times \frac{TP1+TP2}{2}}{2 \times \frac{TP1+TP2}{2} + (N - TP1) + (M - TP2)}. \tag{13}$$

Using the notation of (14), (15) and (16), we can rewrite the $F1^*$ probability as a $F1$-like formula in (17), using as true positive the arithmetic mean of $TP1$ and $TP2$.

$$TP^* = \frac{TP1 + TP2}{2}. \tag{14}$$

$$FN^* = N - TP1. \tag{15}$$

$$FP^* = M - TP2. \tag{16}$$

$$F1^* = \frac{2 \times TP^*}{2 \times TP^* + FN^* + FP^*}. \tag{17}$$

Hence, we obtain that $F1$ is a good approximation to this probability, as we expect $TP1$ to be similar to $TP2$ because, by construction, the differences between them can come only from the cases of multiple overlap and those are infrequent. In fact, if we analyze the MASS SS2 annotations we find 44 cases of multiple overlap among experts' annotations. These are the only cases that could make our novel metric numerically different from the classical $F1$, in a total of 33,458 annotations, which accounts for less than 0.15%. Some authors have even proposed to discard the cases of multiple overlap when computing the $F1$ metric, as it is ill-defined for those cases [15]. More important, in our opinion, the conceptual and methodological differences between the metrics are more significant than the small numerical differences that might arise. The derivation followed to define this probability is symmetric with respect to the row and column indexing, hence, it can be used to determine the degree of agreement between two observations in general, eliminating the need to set one of them as the ground truth. That is of fundamental relevance in a problem with low inter-expert agreement, such as the one we are facing. When the classic $F1$ metric is used for the estimation of inter-expert agreement one of the experts is implicitly granted the condition of ground truth (whether it is done by selection of the experimenter or just randomly), which does not conceptually apply to the problem at hand. Furthermore, to our knowledge, there is no validated conceptual framework in the literature that does not rely on the classical confusion matrix to define TP, TN, FP, and FN as a prior step to calculate the metrics. Even we are forcing the notation to present our contribution in a familiar and understandable way, but it is possible that in the future we will need to move beyond the terms “true” and “false” adjectives to the “positive” and “negative” pairs, as we do have some uncertainty in the reference labels, especially when we are trying to measure that uncertainty. To the best of our knowledge, this is the first time an $F1$-like formula is derived from a probability and not from the harmonic mean of precision and recall.

III. Results

A. Feature selection

As previously introduced, we here make explicit the list of features selected in Table II. The first three columns serve to identify each feature, comprising the type, window length, and band, while the last one represents the number of times it was selected in the bootstrapped feature selection methodology.

TABLE II
Feature selection results.

Type	**Window length (seconds)**	**Band**	**Number of times selected (N_{sel})**
Sigma Index	2	-	1000
Hjort activity	2	Sigma	1000
Sigma Index	1.5	-	1000
Hjort activity	1	Sigma	1000
Sigma index	1	-	1000
Hjort activity	2	Beta1	1000
Hjort activity	2	Beta2	1000
Hjort activity	1.5	Sigma	999
Hjort mobility	2	Beta1	994
Hjort activity	2	Delta2	986
Hjort activity	2	Broadband	985
Hjort mobility	2	Sigma	975
Hjort activity	2	Alpha	971
Hjort activity	1.5	Beta2	918
Sigma index	0.5	-	899
Hjort mobility	2	Beta2	897
Hjort mobility	2	Alpha	893
Hjort mobility	2	Beta1	838
Hjort Mobility	1.5	Beta1	824
Hjort activity	1.5	Beta1	725
Petrosian f. d.	2	Broadband	715
Hjort activity	1	Broadband	680
Hjort activity	2	Theta	652
Hjort activity	0.5	Sigma	650
Hjort mobility	2	Theta	597
Hjort activity	0.5	Alpha	547
Hjort activity	0.5	Delta2	486
Hjort activity	1	Beta1	466
Hjort activity	1	Alpha	389
Hjort activity	1.5	Alpha	383
Hjort mobility	2	Delta2	356
Hjort activity	2	Delta1	354
Hjort activity	1	Theta	351
Hjort complexity	0.5	Broadband	335
Hjort mobility	1.5	Sigma	275
Hjort mobility	1	Beta2	260

B. Optimal Spinco hyperparameters

The hyperparameter optimization in MASS and COGNITION databases is performed, for each fold of the LOOCV procedure [26] and for IoU thresholds of 0.2 and 0.3, using grid search over the validation subjects, as previously explained. In the MASS database the grid values are set to 0.1, 0.2, 0.3, 0.4, 0.5, and 0.6 for the threshold hyperparameter and 10, 20, 30, 40, 50, and 60 for the number of boosting iterations. In the COGNITION database, the grid search values are the same for the threshold parameter but are 5, 10, 15, 20, 30, and 40 for the number of boosting iterations. We experimentally observed that a per-fold hyperparameter optimization led to improved performance compared to using the same values of the hyperparameters for all the folds.

C. Comparison with human-level performance

We propose to make use of the $F1^*$ developed metric in order to estimate the inter-expert agreement using the two experts of the MASS database. Additionally, we estimate the performance of SpinCo against an unseen expert, training and validating SpinCo with each expert and testing against the other in the two possible combinations. Thus, we get a quantification of our system’s inter-expert generalization that can be compared with the inter-expert agreement. The results are shown in Fig. 4. In this experiment, we only considered the spindles detected in N2 as the experts only labeled in this stage [34]. We consider that achieving a model with superior inter-expert generalization compared to inter-expert agreement would result in our system’s surpassing human-level performance in detecting spindle events [12]. On average, SpinCo surpasses the inter-expert agreement for low IoU thresholds, approximately up to 0.35. However, the differences are not statistically significant at a 95% confidence interval for IoU thresholds lower than 0.6, as determined by a paired Wilcoxon test.

D. Spinco benchmarking

In this section, we evaluate the performance of SpinCo against the expert that was trained with, using the metrics previously explained, as shown in Table III. The values in Table III correspond to the test set evaluation. We are reporting the values directly using $F1^*$ as we have conducted experiments comparing those values to classical $F1$ and there are no differences in the first two significant figures in any of the considered cases. This is due to the fact that the differences

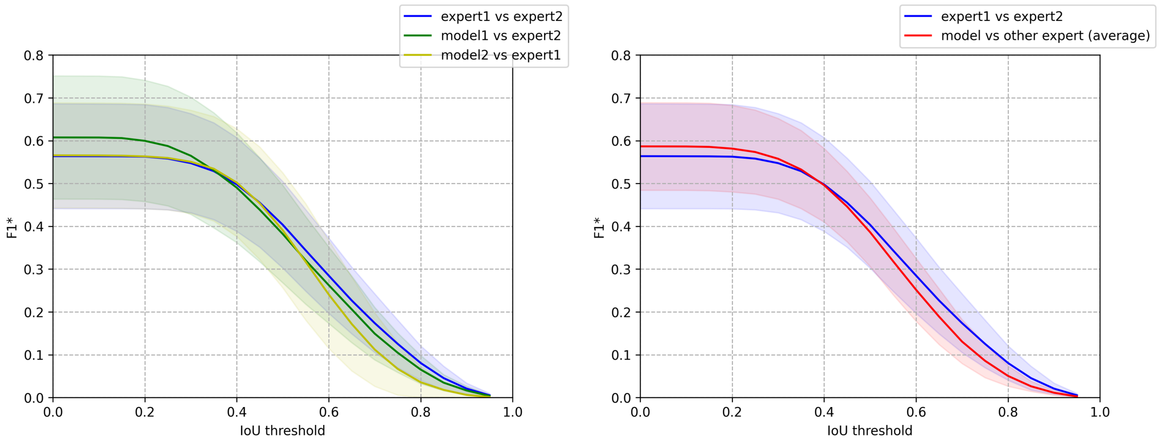

Fig. 4. **Assessment of the method's performance in comparison to expert evaluations.** Left: comparison of the inter-expert agreement, in blue, and the automatic system's inter-expert generalization. In green the model trained and validated with expert 1 compared to expert 2, in yellow the model trained and validated with expert 2 and compared to expert 1. Right: in blue the same inter-expert agreement curve than the left. In red, the average of the yellow and green curves of the left. All shaded regions are by-subject standard deviation calculated over the 15 subjects of MASS SS2 considered.

arise only because of the cases of multiple overlap that are relatively infrequent in real cases, as previously explained.

Comparing the global results with those restricted to N2, we see a different trend in MASS and COGNITION databases. Whereas in the COGNITION database, the results do not change significantly, in the MASS database we see a considerable increase in precision and a consequent slight increase in $F1^{*}$, as recall remains stable.

Table IV shows the correlations and mean differences between per-subject spindle characteristics derived from our approach and those from expert annotations (more details are included in the supplementary materials in our repository [39]). Importantly, all correlations (i.e., R-squared values) between the subject-wise spindle characteristics derived from SpinCo and the expert annotations exceeded 0.6.

E. Performance comparison of SpinCo for different subsets of features

To further see the influence of the most important features for sleep spindle detection, and according to the number of times each feature is selected (N_{sel}) in the bootstrapped procedure, we have considered several thresholds ($N_{sel} = 250$, $N_{sel} = 500$, $N_{sel} = 750$, $N_{sel} = 900$, and $N_{sel} = 1000$) to obtain different subsets of features. Subsequently, SpinCo was trained and evaluated using each of these feature subsets. Table V shows the $F1^{*}$ results of these models. Interestingly, sleep spindles detection performance remains very high using only seven features ($N_{sel} = 1000$ rows).

IV. Discussion

In this study, we have shown the feasibility of extracting signal features using sliding windows at the same sample rate as the original signal. We also have explored the possibility of considering multiple window sizes simultaneously to describe each sample, generalizing our approach further than the classical epoch-by-epoch definition of features in ML systems. When analyzing the feature selection obtained aggregating by window, we see a clear trend of preference for longer windows. In fact, 17 out of 36 features are selected with the longest window proposed of 2 seconds, while only 5 are selected with the shortest one of 0.5 seconds. This can be an indicator of the relevance of the context in which the spindle appears as previously suggested [49]. Furthermore, it indicates a possible direction of improvement for techniques extracting features at 0.5 seconds or shorter [15], [18], [44].

As previously mentioned, our method offers the advantage of being easily interpretable when compared to DL approaches. In this regard, the sigma index was selected at all 4 scales and the majority of the remaining features were Hjort activities and Hjort mobilities. All the frequency bands are selected for the activity parameters: sigma, beta1, beta2, delta1, delta2, theta, alpha, and broadband. This implies that our system is not only detecting spindles present in the sigma band but also rejecting possible positive candidates based on the power simultaneously present in other bands. This suggests that while other detectors have an explicit rejection of alpha waves implemented [47], [50], our method infers it automatically. We hypothesize also that the selection of the broadband activity is linked to the rejection of artifacts in the EEG. Additionally, the mobility parameter, representing the mean can be an indicator of the presence of spindles in the sigma band. Interestingly, this parameter is also selected in the theta, beta1, beta2, alpha, and delta1 bands indicating the particular frequency distribution on these bands can be also a rejection criterion for spindle candidates. Lastly, the complexity of the broadband signal is known to increase in the presence of spindles [49], which justifies the selection of the broadband Petrosian fractal dimension and Hjort Complexity at two different scales. Together, this information provides us with crucial insights not only for the selection of more valuable features in future spindle detectors but also offers pertinent data that could be of potentially valuable for clinical applications.

We also observed that the seven most important EEG

TABLE III
BENCHMARKING RESULTS INCLUDING BOTH THE RESULTS CONSIDERING THE WHOLE SIGNAL AND THOSE RESTRICTED TO N2.

Database	**IoU**	**F1***	**Precision**	**Recall**	**N2 F1***	**N2 Precision**	**N2 Recall**
MASS SS2, E1	0.2	0.724 ± 0.071	0.679 ± 0.142	0.824 ± 0.123	0.752 ± 0.071	0.728 ± 0.151	0.824 ± 0.123
	0.3	0.721 ± 0.070	0.682 ± 0.145	0.814 ± 0.125	0.749 ± 0.069	0.731 ± 0.152	0.814 ± 0.124
MASS SS2, E2	0.2	0.768 ± 0.065	0.726 ± 0.114	0.844 ± 0.108	0.818 ± 0.053	0.814 ± 0.098	0.844 ± 0.108
	0.3	0.755 ± 0.065	0.709 ± 0.107	0.833 ± 0.106	0.805 ± 0.053	0.798 ± 0.096	0.833 ± 0.105
COGNITION	0.2	0.711 ± 0.106	0.720 ± 0.106	0.730 ± 0.174	0.725 ± 0.103	0.728 ± 0.111	0.745 ± 0.164
	0.3	0.699 ± 0.101	0.698 ± 0.100	0.732 ± 0.184	0.714 ± 0.098	0.711 ± 0.107	0.746 ± 0.172

IoU: intersection over union. All values are mean ± std among test subjects.

TABLE IV
CORRELATION BETWEEN EXPERTS AND OUR APPROACH FOR PER-SUBJECT SPINDLE CHARACTERISTICS.

Spindle characteristic (s)	**Database**	**IoU**	**R-squared**	**Mean difference (s)**
Mean spindle duration (s)	MASS SS2, E1	0.2	0.636	-0.071
		0.3	0.639	-0.072
	MASS SS2, E2	0.2	0.631	-0.055
		0.3	0.651	-0.046
	COGNITION	0.2	0.907	-0.105
		0.3	0.956	-0.099
Spindle density (epm)	MASS SS2, E1	0.2	0.635	0.313
		0.3	0.605	0.284
	MASS SS2, E2	0.2	0.723	0.524
		0.3	0.759	0.575
	COGNITION	0.2	0.916	0.058
		0.3	0.920	0.080

epm: events per minute. The correlation is calculated between values determined by expert annotations and those derived from SpinCo. The mean difference is calculated as the average difference between the value estimated by automated detection and the ground truth value provided by expert annotations.

TABLE V
BENCHMARKING RESULTS FOR DIFFERENT SUBSETS OF FEATURES.

Database	**IoU**	N_{sel}	N_f	**F1***	**N2 F1***
MASS SS2, E1	**0.2**	**250**	**36**	**0.724±0.071**	**0.752±0.071**
		500	26	0.719±0.070	0.748±0.070
		750	19	0.722±0.070	0.750±0.070
		900	14	0.723±0.070	0.751±0.069
		1000	7	0.716±0.085	0.748±0.076
	0.3	**250**	**36**	**0.721±0.070**	**0.749±0.069**
		500	26	0.718±0.070	0.746±0.069
		750	19	0.720±0.069	0.748±0.069
		900	14	0.718±0.072	0.746±0.072
		1000	7	0.711±0.083	0.743±0.075
MASS SS2, E2	**0.2**	**250**	**36**	**0.768±0.064**	**0.818±0.053**
		500	26	0.762±0.067	0.815±0.057
		750	19	0.762±0.070	0.812±0.058
		900	14	0.760±0.064	0.810±0.059
		1000	7	0.741±0.079	0.801±0.077
	0.3	**250**	**36**	**0.755±0.065**	**0.805±0.053**
		500	26	0.750±0.070	0.801±0.059
		750	19	0.749±0.070	0.799±0.059
		900	14	0.747±0.068	0.797±0.060
		1000	7	0.728±0.077	0.787±0.078
COGNITION	**0.2**	**250**	**36**	**0.711±0.106**	**0.725±0.103**
		500	26	0.680±0.122	0.692±0.120
		750	19	0.674±0.121	0.688±0.119
		900	14	0.679±0.117	0.694±0.101
		1000	7	0.693±0.068	0.707±0.069
	0.3	**250**	**36**	**0.699±0.101**	**0.714±0.098**
		500	26	0.692±0.104	0.704±0.099
		750	19	0.665±0.113	0.680±0.111
		900	14	0.679±0.098	0.694±0.101
		1000	7	0.689±0.066	0.703±0.067

IoU: intersection over union. N_{sel}: Bootstrap threshold. N_f: Number of features; All values are mean ± std among test subjects.

features in SpinCo play a key role in sleep spindle detection: the sigma index at 1-, 1.5-, and 2-second scales; Hjort activity in the sigma band at 1- and 2-second scales; and Hjort activity in the beta1 and beta2 bands at the 2-second scale. On one hand, the sigma index and sigma band power are well-known features to discriminate sleep spindles [49]. Conversely, beta waves (beta1 and beta2: 13-30 Hz) have been considered for wakefulness [51] and sleep arousal scoring [1], so this parameter can also help to distinguish them from sleep spindles. Overall, we think that this set of seven EEG features could be used to improve sleep spindle scoring rules.

Based on the estimations of inter-expert generalization ability compared to inter-expert agreement, we consider that SpinCo is, at least, as reliable as a human expert using the IoU thresholds of 0.2 and 0.3. Furthermore, this type of test can be useful in the evaluation of automatic spindle detection techniques when few experts have annotated the data. Previous approaches only quantified the ability to reproduce the criterion of a single expert, or a consensus chosen as ground truth, and compared with measures of inter-expert agreement. In our case we can make a fairer comparison as estimations of inter-expert generalization can be obtained based on the $F1^*$ metric developed. It is important to highlight that the inter-expert generalization test would be very difficult to justify using the classical $F1$ approach, as there is no proper ground truth in this case. It is only the availability of the $F1^*$ metric that allows us to consider this type of test. Again, the numerical output would be equivalent but is at the conceptual level where the difference is present.

With respect to the metric values exposed in the benchmarking section, we hypothesize that in MASS we are able to detect spindles in N2 and N3 when training the model with all data available. This leads to an increase in precision when the results are restricted to N2 as spindles are only annotated in this sleep stage. The results of SpinCo in the COGNITION database also support that hypothesis as the metric values do not change significantly when the evaluation is restricted to N2 and this database is annotated in N2 and N3. Overall, this is an advantage compared to other methods trained and evaluated only in N2. Regarding the estimation of subject-wise spindle characteristics, the R-squared values obtained by our approach ($>$ 0.6), in line with the guidelines of Moore & Kirkland [52], indicate a moderate (0.5 $<$ R-squared $<$ 0.7) and/or strong agreement between SpinCo and expert annotations. Similarly, Tapia-Rivas et al. [26] reported high correlations between experts and automatic detectors for subject-level parameters of sleep spindles and k-complexes. However, their results for

TABLE VI
COMPARISON OF THE BENCHMARKING RESULTS WITH OTHER AUTOMATIC SPINDLE DETECTORS IN MASS DATABASE

Spindle detector	**Method**	**Expert**	**N2 F1***	
			IoU>0.2	**IoU>0.3**
MUSSDET [19][a]	ML	E1	-	0.38±0.15[b]
Spinky [16][a]	ML	E1	-	0.45±0.19[b]
McSleep [17][a]	ML	E1	-	0.50±0.20[b]
A7 [15][a]	ML	E1	0.730±0.034[c]	-
		E2	0.749±0.028[c]	-
DOSED [22][a]	DL	E1	0.768±0.029[c]	0.74±0.06[b]
		E2	0.825±0.025[c]	-
SpindleU-Net [25][a]	DL	E1	0.803±0.019[d]	-
		E2	0.854±0.027	-
SEED [26][a]	DL	E1	0.808±0.021	-
		E2	0.861±0.020	-
SpinCo: our proposal	ML	E1	0.752±0.071	0.749±0.069
		E2	0.818±0.053	0.805±0.053

[a]Using classical F1 instead of F1*. [b]Metric values from [22]. [c]Metric values from [26]. [d]Metric values using 19 subjects from MASS database. DL: deep learning. IoU: intersection over union. ML: machine learning. All values are mean ± std among test subjects.

spindle characteristics are not directly comparable as they only used the MODA dataset for this analysis [26].

In Table VI, we compare the results obtained with SpinCo and other state-of-the-art techniques in MASS SS2. Other studies have used different evaluation criteria, for example in [44], $F1$ by-event was estimated close to 0.7 at $IoU > 0.2$ in MASS SS2 with a random under-sampling boosting approach, using the intersection of the annotations of both experts. In our study, our spindle detection framework has been proven effective and is evaluated as the best ML technique, comparable to DL techniques but slightly underperforming in benchmarking metrics, with the advantage of not losing interpretability.

Finally, we have clarified the calculations required to define the by-event metric used in sleep spindle event detection, enhancing its interpretability by defining it in terms of probability. With this new metric, we can fairly measure inter-expert agreement (see Fig. 4). This quantity was similarly estimated in [13] using a private database, where an $F1$ value at $IoU > 0.2$ is reported as 0.61 ± 0.06. To quantify the inter-expert generalization in MASS SS2, we used a leave-one-expert-out on top of our leave-one-subject-out cross-validation schema, similar to what is done in previous studies, as in [27], where the expert agreement with a crowdsourced ground truth was estimated as 0.67 ± 0.07 $F1$ at $IoU > 0.2$. Our value, using a direct comparison of two experts, is comparable but does not depend on the definition of crowdsourced ground truth. In fact, defining a ground truth to apply the $F1^*$ metric is no longer needed, which opens the door to new tests of performance. Our measurements of inter-expert agreement and inter-expert generalization of our automatic system in the MASS SS2 database are statistically equivalent for IoU thresholds lower than 0.6. This implies a solid ground to consider that current state-of-the-art spindle detection algorithms are as reliable as a human expert when applied to the same data population that has been used for training.

One of the main limitations of this type of problem is the dependence on expert annotations as the agreement among experts is far from perfect, but that is unavoidable for a supervised learning task. The use of a crowdsourced ground truth has shown promising results in the evaluation of methods [15], [23], [27]. Nevertheless, the effort necessary to create the annotations makes it difficult to use this approach when creating a model for a specific population, as we aim to do for pediatric patients. That is the reason we evaluated the methodology in terms of the ability to imitate a single expert's criterion, although our performance estimations might be biased toward that particular expert. Similarly, the $F1^*$ metric is more permissive than the classical $F1$ metric in the cases of multiple overlap, which can affect spindle density, although their effect is minimal given that, in most cases, annotation-detection pairs are unique. Another related limitation is the lack of a greater number of complete EEG signals labeled by experts, not only in the N2 stage but in all sleep stages. That would increase the reliability of the detectors in their application to a raw signal, the main use case for these systems. In this respect, a larger database including not only complete EEG signals but also clinical biomarkers (e.g., insomnia or cognitive parameters) would allow us to fully assess the clinical applicability of our proposed approach.

Our measures in MASS of inter-expert generalization of the method and comparison with inter-expert agreement suggest that SpinCo can be safely considered, at least, as accurate as an expert. It can be postulated that obtaining a higher performance in inter-expert generalization than the inter-expert agreement indicates that the automatic detection methodology is successfully capturing systematic aspects of the scoring while not being dramatically impacted by within-expert unreliability. Further studies need to be carried out to determine the causes of the slight drop in performance in the application of SpinCo to the pediatric population compared to adults. We hypothesize that the main causes are the lesser number of signals annotated as well as the fact that annotating sleep spindles in the whole signal is an even more challenging, time-consuming, and, therefore, error-prone task than annotating only in N2 for a human expert.

In our study, we provide a validation of a spindle detection algorithm, SpinCo, but the structure followed prevents us from creating a single model that can be distributed. In the future, we consider the use of databases that rely on a crowdsourced definition of the ground truth based on the annotations of many experts [30] [31] to overcome this limitation.

V. CONCLUSION

We conclude that our approach, SpinCo, is able to imitate the criterion of a given expert successfully by the creation of a different model during training. The benchmarking of automatic sleep spindle detection systems with the same expert used during training have the risk of overfitting to that experts' criterion. A reasonable way to avoid that risk is ensuring inter-expert generalization ability, as demonstrated in this study.

Our approach has been tested in a private database of pediatric patients obtaining comparable results to those of the public adult database. Apart from increasing the robustness of the results, our findings suggest that our system can be used safely to automatically annotate spindles in pediatric patients. Furthermore, benchmarking results are comparable to some DL techniques but come with the advantage of a more interpretable decision-making process as the model depends on a fixed set of features. In this respect, we found a set of seven EEG features playing the main role in sleep spindle detection. Conversely, the feature selection part of the detection framework proposed, based on a feature selection over a feature set that contains a description of different EEG bands, can serve to discover features relevant for detection in other types of microevents. Although, in our study, it has only been tested with spindles.

New developments on the by-event metric used to evaluate spindle detection performance of automatic methods have been proposed. These modifications, firstly, clarify the way the metric is calculated, and secondly, extend its interpretability as it is defined now in terms of probability. Furthermore, the symmetricity in the way the metric is calculated allows us to use it as a similarity measure among two observations in general, without the need to artificially set one as the ground truth. This novelty allows us to measure the similarity between the application of a model trained with annotations of a given expert and the annotations proposed by a different expert, as we do in this study to estimate inter-expert generalization. We propose in this work that the comparison of this value with direct estimations of inter-expert agreement can be considered a test of whether an automatic system reaches human-level performance or not. Future studies could use this metric, $F1^*$, in other signal detection tasks with low inter-expert agreement or extend it to image detection problems.

VI. References

[1] R. B. Berry, R. Budhiraja, D. J. Gottlieb, D. Gozal, C. Iber, V. K. Kapur, C. L. Marcus, R. Mehra, S. Parthasarathy, S. F. Quan, S. Redline, K. P. Strohl, S. L. D. Ward, and M. M. Tangredi, "Rules for Scoring Respiratory Events in Sleep: Update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events," *Journal of Clinical Sleep Medicine*, vol. 08, no. 05, pp. 597–619, Oct. 2012, publisher: American Academy of Sleep Medicine. [Online]. Available: https://jcsm.aasm.org/doi/full/10.5664/jcsm.2172

[2] C. Iber, "The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications," 2007.

[3] S. M. Purcell, D. S. Manoach, C. Demanuele, B. E. Cade, S. Mariani, R. Cox, G. Panagiotaropoulou, R. Saxena, J. Q. Pan, J. W. Smoller, S. Redline, and R. Stickgold, "Characterizing sleep spindles in 11,630 individuals from the National Sleep Research Resource," *Nature Communications*, vol. 8, no. 1, p. 15930, Jun. 2017, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/ncomms15930

[4] T. T. Dang-Vu, A. Salimi, S. Boucetta, K. Wenzel, J. O'Byrne, M. Brandewinder, C. Berthomier, and J.-P. Gouin, "Sleep Spindles Predict Stress-Related Increases in Sleep Disturbances," *Frontiers in Human Neuroscience*, vol. 9, p. 68, 2015. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2015.00068

[5] S. M. Fogel and C. T. Smith, "The function of the sleep spindle: A physiological index of intelligence and a mechanism for sleep-dependent memory consolidation," *Neuroscience & Biobehavioral Reviews*, vol. 35, no. 5, pp. 1154–1165, Apr. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0149763410002046

[6] Z. Fang, V. Sergeeva, L. B. Ray, J. Viczko, A. M. Owen, and S. M. Fogel, "Sleep Spindles and Intellectual Ability: Epiphenomenon or Directly Related?" *Journal of Cognitive Neuroscience*, vol. 29, no. 1, pp. 167–182, Jan. 2017. [Online]. Available: https://doi.org/10.1162/jocn_a_01034

[7] Z. Clemens, D. Fabó, and P. Halász, "Overnight verbal memory retention correlates with the number of sleep spindles," *Neuroscience*, vol. 132, no. 2, pp. 529–535, Jan. 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306452205000941

[8] E. Huupponen, S. L. Himanen, J. Hasan, and A. Värri, "Automatic analysis of electro-encephalogram sleep spindle frequency throughout the night," *Medical and Biological Engineering and Computing*, vol. 41, no. 6, pp. 727–732, Nov. 2003. [Online]. Available: https://doi.org/10.1007/BF02349981

[9] Y.-Y. Weng, X. Lei, and J. Yu, "Sleep spindle abnormalities related to Alzheimer's disease: a systematic mini-review," *Sleep Medicine*, vol. 75, pp. 37–44, Nov. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389945720303609

[10] E. J. Wamsley, M. A. Tucker, A. K. Shinn, K. E. Ono, S. K. McKinley, A. V. Ely, D. C. Goff, R. Stickgold, and D. S. Manoach, "Reduced Sleep Spindles and Spindle Coherence in Schizophrenia: Mechanisms of Impaired Memory Consolidation?" *Biological Psychiatry*, vol. 71, no. 2, pp. 154–161, Jan. 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006322311008146

[11] L. De Gennaro and M. Ferrara, "Sleep spindles: an overview," *Sleep Medicine Reviews*, vol. 7, no. 5, pp. 423–440, Oct. 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1087079202902522

[12] L. Ray, S. Sockeel, M. Soon, A. Bore, A. Myhr, B. Stojanoski, R. Cusack, A. Owen, J. Doyon, and S. Fogel, "Expert and crowd-sourced validation of an individualized sleep spindle detection method employing complex demodulation and individualized normalization," *Frontiers in Human Neuroscience*, vol. 9, p. 507, 2015. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2015.00507

[13] S. L. Wendt, P. Welinder, H. B. D. Sorensen, P. E. Peppard, P. Jennum, P. Perona, E. Mignot, and S. C. Warby, "Inter-expert and intra-expert reliability in sleep spindle scoring," *Clinical Neurophysiology*, vol. 126, no. 8, pp. 1548–1556, Aug. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245714007329

[14] S. L. Wendt, J. A. E. Christensen, J. Kempfner, H. L. Leonthin, P. Jennum, and H. B. D. Sorensen, "Validation of a novel automatic sleep spindle detector with high performance during sleep in middle aged subjects," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2012, pp. 4250–4253, iSSN: 1558-4615.

[15] K. Lacourse, J. Delfrate, J. Beaudry, P. Peppard, and S. C. Warby, "A sleep spindle detection algorithm that emulates human expert spindle scoring," *Journal of Neuroscience Methods*, vol. 316, pp. 3–11, Mar. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027018302504

[16] T. Lajnef, C. O'Reilly, E. Combrisson, S. Chaibi, J.-B. Eichenlaub, P. M. Ruby, P.-E. Aguera, M. Samet, A. Kachouri, S. Frenette, J. Carrier, and K. Jerbi, "Meet Spinky: An Open-Source Spindle and K-Complex Detection Toolbox Validated on the Open-Access Montreal Archive of Sleep Studies (MASS)," *Frontiers in Neuroinformatics*, vol. 11, p. 15, 2017.

[17] A. Parekh, I. W. Selesnick, R. S. Osorio, A. W. Varga, D. M. Rapoport, and I. Ayappa, "Multichannel sleep spindle detection using sparse low-rank optimization," *Journal of Neuroscience Methods*, vol. 288, pp. 1–16, Aug. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027017301681

[18] C. R. Patti, S. S. Shahrbabaki, C. Dissanayaka, and D. Cvetkovic, "Application of random forest classifier for automatic sleep spindle detection," in *2015 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Oct. 2015, pp. 1–4.

[19] D. Lachner-Piza, N. Epitashvili, A. Schulze-Bonhage, T. Stieglitz, J. Jacobs, and M. Dümpelmann, "A single channel sleep-spindle detector based on multivariate classification of EEG epochs: MUSSDET," *Journal of Neuroscience Methods*, vol. 297, pp. 31–43, Mar. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027017304417

[20] L. Wei, S. Ventura, S. Mathieson, G. B. Boylan, M. Lowery, and C. Mooney, "Spindle-AI: Sleep Spindle Number and Duration Estimation in Infant EEG," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 1, pp. 465–474, Jan. 2022, conference Name: IEEE Transactions on Biomedical Engineering. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9490357

[21] X. Zhuang, Y. Li, and N. Peng, "Enhanced automatic sleep spindle detection: a sliding window-based wavelet analysis and comparison using a proposal assessment method," *Applied Informatics*, vol. 3, no. 1, p. 11, Dec. 2016. [Online]. Available: https://doi.org/10.1186/s40535-016-0027-9

[22] S. Chambon, V. Thorey, P. J. Arnal, E. Mignot, and A. Gramfort, "DOSED: A deep learning approach to detect multiple sleep micro-events in EEG signal," *Journal of Neuroscience Methods*, vol. 321, pp. 64–78, Jun. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027019301013

[23] L. Kaulen, J. T. C. Schwabedal, J. Schneider, P. Ritter, and S. Bialonski, "Advanced sleep spindle identification with neural networks," *Scientific Reports*, vol. 12, no. 1, p. 7686, May 2022, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-022-11210-y

[24] P. M. Kulkarni, Z. Xiao, E. J. Robinson, A. S. Jami, J. Zhang, H. Zhou, S. E. Henin, A. A. Liu, R. S. Osorio, J. Wang, and Z. Chen, "A deep learning approach for real-time detection of sleep spindles," *Journal of Neural Engineering*, vol. 16, no. 3, p. 036004, Mar. 2019, publisher: IOP Publishing. [Online]. Available: https://doi.org/10.1088/1741-2552/ab0933

[25] J. You, D. Jiang, Y. Ma, and Y. Wang, "SpindleU-Net: An Adaptive U-Net Framework for Sleep Spindle Detection in Single-Channel EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1614–1623, 2021, conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.

[26] N. I. Tapia-Rivas, P. A. Estévez, and J. A. Cortes-Briones, "A robust deep learning detector for sleep spindles and K-complexes: towards population norms," *Scientific Reports*, vol. 14, no. 1, p. 263, jan 2024. [Online]. Available: https://www.nature.com/articles/s41598-023-50736-7

[27] S. C. Warby, S. L. Wendt, P. Welinder, E. G. S. Munk, O. Carrillo, H. B. D. Sorensen, P. Jennum, P. E. Peppard, P. Perona, and E. Mignot, "Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods," *Nature Methods*, vol. 11, no. 4, pp. 385–392, Apr. 2014, number: 4 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nmeth.2855

[28] C. O'Reilly, S. C. Warby, and T. Nielsen, "Editorial: Sleep Spindles: Breaking the Methodological Wall," *Frontiers in Human Neuroscience*, vol. 10, p. 672, 2016.

[29] C. O'Reilly and T. Nielsen, "Automatic sleep spindle detection: benchmarking with fine temporal resolution using open science tools," *Frontiers in Human Neuroscience*, vol. 9, 2015. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2015.00353

[30] B. D. Yetton, K. Lacourse, J. Delfrate, S. Mednick, and S. Warby, "The MODA sleep spindle dataset: A large, open, high quality dataset of annotated sleep spindles," May 2016, publisher: OSF. [Online]. Available: https://osf.io/8bma7/

[31] K. Lacourse, B. Yetton, S. Mednick, and S. C. Warby, "Massive online data annotation, crowdsourcing to generate high quality sleep spindle annotations from EEG data," *Scientific Data*, vol. 7, no. 1, p. 190, Jun. 2020, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41597-020-0533-4

[32] M. Borsky, M. Serwatko, E. S. Arnardottir, and J. Mallett, "Toward Sleep Study Automation: Detection Evaluation of Respiratory-Related Events," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 3418–3426, Jul. 2022, conference Name: IEEE Journal of Biomedical and Health Informatics. [Online]. Available: https://ieeexplore.ieee.org/document/9736681

[33] K. Campbell, A. Kumar, and W. Hofman, "Human and automatic validation of a phase-locked loop spindle detection system," *Electroencephalography and Clinical Neurophysiology*, vol. 48, no. 5, pp. 602–605, May 1980. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0013469480902965

[34] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research," *Journal of Sleep Research*, vol. 23, no. 6, pp. 628–635, 2014, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.12169. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.12169

[35] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[36] S. Liu, B. Fu, W. Wang, M. Liu, and X. Sun, "Dynamic Sepsis Prediction for Intensive Care Unit Patients Using XGBoost-Based Model With Novel Time-Dependent Features," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4258–4269, Aug. 2022, conference Name: IEEE Journal of Biomedical and Health Informatics. [Online]. Available: https://ieeexplore.ieee.org/document/9767649

[37] H. Shin, "XGBoost Regression of the Most Significant Photoplethysmogram Features for Assessing Vascular Aging," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 3354–3361, Jul. 2022, conference Name: IEEE Journal of Biomedical and Health Informatics. [Online]. Available: https://ieeexplore.ieee.org/document/9713685

[38] S. Devuyst, "The DREAMS Databases and Assessment Algorithm," Jan. 2005. [Online]. Available: https://zenodo.org/record/2650142

[39] E. Gurdiel, "Python code of the publication "Beyond the ground truth, XGBoost model applied to sleep spindle event detection"." 2023. [Online]. Available: https://github.com/quiquegurdiel/SpinCo

[40] G. C. Gutiérrez-Tobal, J. Gomez-Pilar, L. Kheirandish-Gozal, A. Martín-Montero, J. Poza, D. Álvarez, F. del Campo, D. Gozal, and R. Hornero, "Pediatric Sleep Apnea: The Overnight Electroencephalogram as a Phenotypic Biomarker," *Frontiers in Neuroscience*, vol. 15, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2021.644697

[41] C. R. Patti, T. Penzel, and D. Cvetkovic, "Sleep spindle detection using multivariate Gaussian mixture models," *Journal of Sleep Research*, vol. 27, no. 4, p. e12614, 2018, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.12614. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.12614

[42] E. Huupponen, G. Gómez-Herrero, A. Saastamoinen, A. Värri, J. Hasan, and S.-L. Himanen, "Development and comparison of four sleep spindle detection methods," *Artificial Intelligence in Medicine*, vol. 40, no. 3, pp. 157–170, Jul. 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365707000516

[43] A. Nonclercq, C. Urbain, D. Verheulpen, C. Decaestecker, P. Van Bogaert, and P. Peigneux, "Sleep spindle detection through amplitude–frequency normal modelling," *Journal of Neuroscience Methods*, vol. 214, no. 2, pp. 192–203, Apr. 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027013000423

[44] T. Kinoshita, K. Fujiwara, M. Kano, K. Ogawa, Y. Sumi, M. Matsuo, and H. Kadotani, "Sleep Spindle Detection Using RUSBoost and Synchrosqueezed Wavelet Transform," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 2, pp. 390–398, Feb. 2020, conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.

[45] A. Petrosian, "Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns," in *Proceedings Eighth IEEE Symposium on Computer-Based Medical Systems*, Jun. 1995, pp. 212–217.

[46] J. LaRocco, P. J. Franaszczuk, S. Kerick, and K. Robbins, "Spindler: a framework for parametric analysis and detection of spindles in EEG with application to sleep spindles," *Journal of Neural Engineering*, vol. 15, no. 6, p. 066015, Sep. 2018, publisher: IOP Publishing. [Online]. Available: https://doi.org/10.1088/1741-2552/aadc1c

[47] M. M. Kabir, R. Tafreshi, D. B. Boivin, and N. Haddad, "Enhanced automated sleep spindle detection algorithm based on synchrosqueezing," *Medical & Biological Engineering & Computing*, vol. 53, no. 7, pp. 635–644, Jul. 2015. [Online]. Available: https://doi.org/10.1007/s11517-015-1265-z

[48] E. M. Ventouras, E. A. Monoyiou, P. Y. Ktonas, T. Paparrigopoulos, D. G. Dikeos, N. K. Uzunoglu, and C. R. Soldatos, "Sleep spindle detection using artificial neural networks trained with filtered time-domain EEG: A feasibility study," *Computer Methods and Programs in Biomedicine*, vol. 78, no. 3, pp. 191–207, Jun. 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016926070500060X

[49] J. Gomez-Pilar, G. C. Gutiérrez-Tobal, J. Poza, S. Fogel, J. Doyon, G. Northoff, and R. Hornero, "Spectral and temporal characterization of sleep spindles—methodological implications," *Journal of Neural Engineering*, vol. 18, no. 3, p. 036014, Mar. 2021, publisher: IOP Publishing. [Online]. Available: https://doi.org/10.1088/1741-2552/abe8ad

[50] M. Adamczyk, L. Genzel, M. Dresler, A. Steiger, and E. Friess, "Automatic Sleep Spindle Detection and Genetic Influence Estimation Using Continuous Wavelet Transform," *Frontiers in Human Neuroscience*, vol. 9, 2015. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2015.00624

[51] F. Vaquerizo-Villar, G. C. Gutiérrez-Tobal, E. Calvo, D. Álvarez, L. Kheirandish-Gozal, F. del Campo, D. Gozal, and R. Hornero, "An explainable deep-learning model to stage sleep states in children and propose novel EEG-related patterns in sleep apnea," *Computers in Biology and Medicine*, vol. 165, no. October, p. 107419, 2023.

[52] D. S. Moore and S. Kirkland, *The basic practice of statistics*. WH Freeman New York, 2007, vol. 2.