Scuola universitaria professionale
della Svizzera italiana

SUPSI

University of Applied Sciences and Arts of Southern Switzerland
Department of Innovative Technologies

Applied Case Studies of Machine Learning and Deep Learning in Key Areas II

PERSONALIZED SLEEP SPINDLE DETECTION

Dyuman Bulloni
dyuman.bulloni@student.supsi.ch

Amos Colombo
amos.colombo@student.supsi.ch

Christian Pala
christian.pala@student.supsi.ch

Professor: Francesca D. Faraci
SUPSI, Lugano Switzerland

14/05/2023

Table of Contents

List of Tables

1. Introduction

Sleep spindles are an important electroencephalography (EEG) pattern observed during non-rapid-eye-movement (NREM) sleep stages, with a frequency range of 11 to 16 Hz and a duration of at least 0.5 seconds[1]. They are considered to play a crucial role in sleep-related cerebral plasticity and are believed to mediate many sleep-related functions, from memory consolidation to cortical development [2]. Automated sleep spindle detection algorithms have been developed and show good performance, but their performance deteriorates when applied to different datasets [3].

The automated spindle detection procedure typically involves pre-processing, feature extraction, feature selection, and spindle recognition with a classifier. We follow the approach from [4], to explore if a more personalized algorithm may be beneficial for spindle detection.

In this paper we try to build upon the previous work done in [4] and test various changes to update and validate their results.

2. Data

The data for this study was sourced from the DREAMS Sleep Spindles Database, a widely recognized and publicly accessible resource for sleep studies. This is the same dataset utilized in the previous study [4]. We followed identical signal processing steps as those outlined in [4]. In terms of label generation, we adopted a comprehensive approach. Initially, we employed the scoring provided with the database, as in [4]. Subsequently, we supplemented this procedure by generating new labels using the Yet Another Spindle Algorithm (YASA) [5]. Finally, we combined these labels to create a more robust and comprehensive set for analysis.

2.1 Critique of the Data

The dataset employed in the original paper, while considered standard in the research community and selected for benchmarking purposes, presents several limitations upon further scrutiny. A notable concern is the limited number of patients included in the dataset. This restricts the extent to which meaningful conclusions can be drawn regarding model performance.

Furthermore, the evaluation of this dataset was partially carried out by only two experts, covering different patients and time windows [4].

In contrast, YASA was built on a more diverse dataset evaluated by five different experts [5], achieving state-of-the-art performance in spindle detection.

We advocate for new studies on personalized spindle detection using more robust datasets.

3. Feature extraction

We extracted the same features as in [4], following the same windowed methodology, additionally, we integrated the features present in [5].

- Hjorth parameters of mobility and complexity: The Hjorth parameters of mobility and complexity are measures of the temporal dynamics of the signal. The mobility parameter is a measure of the rate at which the amplitude of the signal changes over time, while the complexity parameter is a measure of the irregularity or complexity of the signal. In the context of EEG signals, the Hjorth parameters can be used as indicators of the dynamics of the underlying neural activity.

- hypnogram: A feature that provides a detailed representation of an individual's current sleeping phase, helping to determine whether it corresponds to a spindle or not.

These parameters have been found useful in characterizing the spectral properties of sleep spindles, which are known to be associated with cognitive processes such as memory consolidation and learning [5].

3.1 Feature ranking

We verified the results of [4] by ranking the features using Maximum Relevance — Minimum Redundancy feature selection (MRMRF). The feature importance we derive is similar but not identical to [4], as shown in Table 1. We also integrated the ranking using Mutual Information (MI) as a different criterion. Our findings suggest that the addition of Hjorth mobility and complexity as features are valuable for the analysis.

Table 1: Feature rankings with MI and MRMRF

MI Ranking		**MRMRF Ranking**	
Rank	**Feature**	**Rank**	**Feature**
1	Phase-amplitude coupling	1	Complexity
2	Power peak	2	Sample entropy
3	Energy ratio	3	Variance
4	Power ratio	4	Power ratio
5	Complexity	5	Hypnogram
6	Zero-crossing rate	6	Mobility
7	Hypnogram	7	Zero-crossing rate
8	Mobility	8	Inter-quartile range
9	Sample entropy	9	Power peak
10	Maximum value	10	Skewness
11	Kurtosis	11	Kurtosis
12	Mean frequency	12	Energy ratio
13	Inter-quartile range	13	Mean frequency
14	Standard deviation	14	Phase-amplitude coupling
15	Variance	15	Standard deviation
16	Minimum value	16	Minimum value
17	Skewness	17	Maximum value

Feature selection often comes with the risk of information loss. Although it can improve model interpretability and computational efficiency, it may inadvertently remove relevant information, which could negatively impact model performance. Considering the complexity and multifaceted nature of sleep spindle detection, it was deemed more prudent to retain the full feature set to encapsulate as many aspects of the data as possible.

This was also motivated by the differences in the rankings produced by the MRMRF and MI methods, as shown in Table 1, which indicated that there was no clear consensus on the most important features. Consequently, the elimination of any feature could potentially result in the loss of valuable information, justifying our decision to retain the complete feature set in this study.

4. Label generation

The dataset comes with spindle detection annotations, provided by 2 experts as well as an automatically generated one, the details are presented in [4]. Following [4] We take the union of the detected spindles from all sources as our ground truth. To this, we supplemented the spindles detected by the YASA algorithm. Our labeling approach was constructed such that a window was designated as containing a spindle if any of the labeling methods at our disposal identified a positive value at any point within the given time interval. Following the procedure described in [4], we eliminated an instance where a window marked with a positive spindle detection was surrounded by windows of negative detection, considering it an outlier. We adopted this labeling methodology to circumvent the issue of scarce positive samples in the dataset, under the assumption that a sample was indeed a spindle if at least one of our sources labeled it as such. In a setting where more data is available, we would opt for a more rigorous methodology to consolidate the various sources.

5. Modeling

In the original paper [4], the authors employed a Support Vector Classifier (SVC) for their analysis. In our study, we expanded upon this by not only implementing the SVC but also exploring additional machine-learning models. The models chosen for this study include the SVC, K-Nearest Neighbors (KNN), Random Forest (RF), and Gradient Boosting (GB). Briefly, these models are characterized as follows:

- **SVC**: A versatile algorithm widely used for its efficacy in classification tasks with clear class separations.
- **KNN**: An intuitive model that performs well in datasets with fewer features, particularly when data points cluster within classes.
- **RF**: An ensemble algorithm adept at handling high-dimensional and noisy datasets, offering robustness against overfitting.
- **GB**: An ensemble model known for its effectiveness in binary classification tasks and capability to handle complex datasets with numerous features.

By applying these diverse models, our aim is to explore a range of methodologies and assess their applicability to our sleep spindle detection task.

5.1 Training, Validation, and Testing

In the referenced paper [4], the authors detail a training and testing procedure to compare personalized and general models. We identified several issues with this approach, particularly:

1. The choice of a very small dataset, 30 observations, for both the global model and individual patients.
2. The random selection of 15 positive and 15 negative instances potentially introduces selection bias and leads to opaque results.
3. Testing on the entire signal after sampling instances for training can result in overoptimistic performance estimates.

To address these concerns, we adopted a more conventional approach to data partitioning. For individual models, we split the data using an 80:20 ratio for training and testing, respectively. For the global model, we randomly select 5 patients for the training set, 1 for validation, and 2 for testing. This strategy prevents the model from overfitting to individual-specific patterns and ensures it learns to recognize a more general spindle pattern. This is particularly challenging, given that the DREAMS database only includes patients with sleep anomalies, lacking a "normal" baseline.

5.1.1 Normalization

We employed the Standard Scaler from Scikit-learn [6] to normalize the data. This normalization process was integrated into a single pipeline for each model to streamline the procedure and ensure consistency across different models.

5.1.2 Evaluation Metrics

Given the significant class imbalance inherent to our dataset, and our prioritization of the positive class (spindle), we adopted two distinct metrics for assessing our models:

- **Binary F-1 Score**: This metric computes the F-1 score solely for the positive class. It is particularly valuable in our case as we are primarily concerned with correctly identifying sleep spindles.
- **Macro F-1 Score**: This metric calculates the unweighted average of the F-1 scores for each class. It provides a holistic measure of the model's performance across all classes without considering class unbalances.

5.2 Data Augmentation

Given the scarcity of positive samples in the database, the preliminary unbalanced models struggled to generalize and often disregarded the minority class. To rectify this, we experimented with various strategies using the imbalanced-learn [7] library:

- **Random Under-Sampling (RUS):** Randomly eliminates samples from the majority class to harmonize the class distribution. Although straightforward, it risks discarding potentially important samples, leading to loss of information.
- **Random Over-Sampling (ROS):** Duplicates random samples from the minority class to balance the class distribution. While beneficial when the minority class has relatively few samples, it can lead to overfitting.
- **SMOTE (Synthetic Minority Over-sampling Technique):** Generates synthetic samples for the minority class based on existing samples' interpolation. It can be effective when the minority class is underrepresented and aids in preventing overfitting.
- **SVM SMOTE:** A variant of SMOTE that employs a Support Vector Machine (SVM) to identify the optimal samples for generating synthetic instances. It can outperform conventional SMOTE when the minority class significantly overlaps with the majority class.
- **ADASYN (Adaptive Synthetic Sampling):** Produces more synthetic samples for the minority class near the decision boundary and fewer samples farther from the boundary. It can be beneficial when the decision boundary between classes is complex and challenging to model.

5.3 Post processing

Adhering to the methodology outlined in [4], we implemented a corrective step to handle potential outliers predicted by the models. Specifically, we ensured that no spindle window was predicted in isolation.

5.4 Global models

Table 2 presents the performance metrics of various algorithms on the global dataset. Mirroring the findings of [4], our models also achieve the best spindle detection results with the Support Vector Classifier (SVC). However, contrary to the results reported in the referenced paper, our scores significantly lag behind the current state-of-the-art benchmarks.

Unbalanced			
Model	**Binary F-1**	**Macro F-1**	**Weighted F-1**
SVC	0.181	0.577	0.930
KNN	0.187	0.579	0.929
RF	0.202	0.588	0.932
GB	0.250	0.612	0.935
Balanced SVC			
Sampler	**Binary F-1**	**Macro F-1**	**Weighted F-1**
Undersampler	**0.291**	0.594	0.863
Oversampler	0.279	0.581	0.851
Balanced GB			
Sampler	**Binary F-1**	**Macro F-1**	**Weighted F-1**
Undersampler	0.212	0.552	0.856
Oversampler	0.211	0.560	0.871

Table 2: Global model score table

6. Hyper-parameter tuning

Hyper-parameter tuning was not mentioned in the original paper[4]. We used the Optuna framework [8] to improve the scores obtained by our baseline models. For some patients, and on the global model, an improvement of 10-45% on the baseline binary F1-score was achieved.

7. Results

To compare the global and personalized approaches, we report our findings using both methodologies.

7.1 Personalized models

The best combination of models and samplers for each patient is shown in Table

Table 3: Best sampler and model combination for each patient

Patient	Sampler	Model	Macro F-1	Binary F-1	Tuned Binary F-1
1	ADASYN	SVC	0.699	0.415	0.458
2	ADASYN	GB	0.589	0.317	0.211
3	SMOTE	SVC	0.676	0.317	0.372
4	SMOTE	SVC	0.698	0.292	0.424
5	RUS	RF	0.513	0.361	0.088
6	ADASYN	KNN	0.615	0.301	0.270
7	SVM SMOTE	KNN	0.492	0.187	0.000
8	RUS	GB	0.658	0.330	0.365

7.2 Global models

In Table 4 we report the results of the global SVC model with the original features from [4] and the new set we developed for this paper.

Features	Binary F-1	Macro F-1	Weighted F-1
original	0.053	0.513	0.922
augmented	0.181	0.577	**0.930**
fine-tuned	**0.322**	**0.629**	0.902

Table 4: Results comparison with original features and fine-tuning

7.3 Discussion

Overall we did not detect a strong improvement by personalizing the algorithm for a specific patient, compared to using a global model. The high degree of variability observed in the results further underscores the need for a more comprehensive analysis. Nonetheless, there are indications of promise also for the personalized approach; particularly in the case of Patient 1 where we had more positive labels, we were able to improve the detection significantly compared to the global baseline.

8. Conclusions

Our study indicates that while personalized models hold promise, they do not necessarily outperform global models in spindle detection tasks. This underscores the importance of a balanced approach, leveraging both global and individual patient data as well as different classification strategies.

References

[1] S. Foundation, *Sleep spindles*, https://www.sleepfoundation.org/how-sleep-works/sleep-spindles, accessed 2023.

[2] T. Andrillon, Y. Nir, R. J. Staba, *et al.*, "Sleep spindles in humans: Insights from intracranial eeg and unit recordings," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 31, no. 49, pp. 17 821–17 834, 2011. DOI: 10.1523/JNEUROSCI.2604-11.2011. [Online]. Available: https://doi.org/10.1523/JNEUROSCI.2604-11.2011.

[3] S. C. Warby, S. L. Wendt, P. Welinder, *et al.*, "Sleep-spindle detection: Crowdsourcing and evaluating performance of experts, non-experts and automated methods," *Nature Methods*, vol. 11, no. 4, pp. 385–392, 2014. DOI: 10.1038/nmeth.2855. [Online]. Available: https://doi.org/10.1038/nmeth.2855.

[4] S. Scafa, L. Fiorillo, M. Lucchini, *et al.*, "Personalized sleep spindle detection in whole night polysomnography," *Frontiers in Neuroscience*, vol. 15, p. 638 232, Jul. 2020. DOI: 10.3389/fnins.2021.638232.

[5] R. Vallat and M. P. Walker, "A universal, open-source, high-performance tool for automated sleep staging," *bioRxiv*, 2021. DOI: 10.1101/2021.05.28.446165. eprint: https://www.biorxiv.org/content/early/2021/05/28/2021.05.28.446165.full.pdf. [Online]. Available: https://www.biorxiv.org/content/early/2021/05/28/2021.05.28.446165.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[7] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, *et al.*, "Personalized spindle detection-Python," *PeerJ*, vol. 2, e453, Jun. 2014, ISSN: 2167-8359. DOI: 10.7717/peerj.453. [Online]. Available: https://doi.org/10.7717/peerj.453.

[8] T. Akiba, S. Sano, T. Yanase, and T. Ohta, *Optuna: A next-generation hyperparameter optimization framework*, https://optuna.org/, accessed 2023.