A Wavelet- and Machine Learning-Based Framework for the Automatic Detection of Artefacts in Electromyography REM sleep

A Wavelet- and Machine Learning-Based Framework for the Automatic Detection of Artefacts in Electromyography REM sleep

Irene Rechichi[1,*], Ambra Stefani[2], Birgit Högl[2], Anna Heidbreder[2,3], Evi Holzknecht[2], Melanie Bergmann[2,3], Abubaker Ibrahim[2], Elisabeth Brandauer[2], Gabriella Olmo[1,‡], and Matteo Cesari[2,‡]

***Abstract*— Rapid eye movement sleep Without Atonia (RWA) is the polysomnographic hallmark of REM Sleep Behavior Disorder (RBD), manifesting either as elevated background tone or phasic activity. Scoring RWA relies heavily on visual inspection of electromyography (EMG) signals from polysomnography (PSG) recordings. This process is time-consuming and prone to inter-rater variability, particularly due to the presence of artefacts. Currently, no standardized method for artefacts removal is available. This study proposes a Matched-Wavelet approach to characterize the morphology of EMG signals, and a Machine Learning (ML) based framework to identify artefacts from EMG recordings during REM sleep, to facilitate subsequent RWA scoring, by decreasing manual labour. The best models achieved F1 scores of 79.2% and 86.3% in detecting artefacts from background and phasic activity, respectively. These results suggest the feasibility of automatically remove artefacts through a low-computational cost method, leading to improved reliability in RWA assessments.**

***Clinical relevance*— The framework provides a robust tool for the assessment of artefacts in EMG recordings, improving the reliability of RWA assessment, and contributing to improved diagnostic accuracy.**

I. INTRODUCTION

Rapid eye movement (REM) sleep behavior disorder (RBD) is a parasomnia characterized by vivid dreams and loss of physiological muscle atonia in REM sleep (REM sleep without atonia - RWA) [1]. Abnormal muscle activity in REM sleep appears either as sustained muscle tone or periodic, rapid bursts of muscular activity [2]. These two are defined as elevated background (also defined tonic) and phasic activity, respectively [3]. Currently, RWA scoring guidelines are regulated by the American Academy of Sleep Medicine (AASM) Manual for Scoring of Sleep and Associated Events [4], and rely primarily on visual inspection and manual scoring of electromyography (EMG) traces from polysomnography (PSG) recordings, in the mentalis, Flexor Digitorum Superficialis (FDS), and Tibialis Anterior (TA) muscles.

Few (semi-)automatic algorithms for scoring RWA have been proposed [5]; of them, only one is integrated into a clinical PSG software [6]. However, its use is still time-consuming, due to the need to manually identify artefacts [7]. Artefacts may be caused by arousals, snoring/respiration (mainly in the mentalis muscle), or technical issues occurring during PSG. A previous study highlighted the need for artefact correction to ensure the robustness of RWA assessment [7]. A significant share of artefacts lies in respiration/snoring, that substantially alters the mentalis EMG activity. As previously specified, artefact removal is handled manually by neurologists or sleep technologists during visual inspection of PSG, which requires extensive and protracted manual effort, and reveals the labor-intensive nature of RWA scoring. Moreover, the lack of a systematic approach in the identification of artefacts leads to significant interrater and intra-rater variability, as stated in a previous study [7].

These open challenges suggest the need for automated artefact detection methods to (1) facilitate the RWA scoring process by reducing manual labor and (2) reduce interrater variability. This study presents a framework for the automatic detection of artefacts in EMG recordings during REM sleep through objective metrics and a low-computational cost approach, to promote the development of fully automated pipelines in the assessment of RWA and improve the reliability of automatic scoring methods.

II. MATERIALS

A. Subjects

This study included 25 subjects (14 males, aged 57.2 $\pm$ 14.9 years), who underwent video-PSG at the Sleep Center, Department of Neurology, Medical University of Innsbruck (Austria). During PSG, mentalis, bilateral FDS and TA EMG signals were recorded with sampling frequency of 1 kHz. Sleep stages were scored according to the AASM criteria [4]. The participants included 8 subjects with RBD; the remainder were randomly included from a cohort under study for suspected parasomnias [7]. This retrospective study was approved by the Ethical Committee of the Medical University of Innsbruck.

B. Data

For each subject, REM sleep was divided into 3-s mini-epochs; a total of 956 $\pm$ 70 mini-epochs across all subjects were included. The previously validated semi-automatic software was used to identify, for each mini-epoch, the presence of either phasic or elevated background activity [6]. Four expert scorers independently performed manual artefact correction on all the activity identified by the software;

*Corresponding author. irene.rechichi@polito.it

[1]Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy

[2]Department of Neurology, Medical University of Innsbruck, 6020 Innsbruck, Austria

[3]Department of Neurology, Johannes Kepler University Linz, 4040 Linz, Austria

‡Shared last authorship.

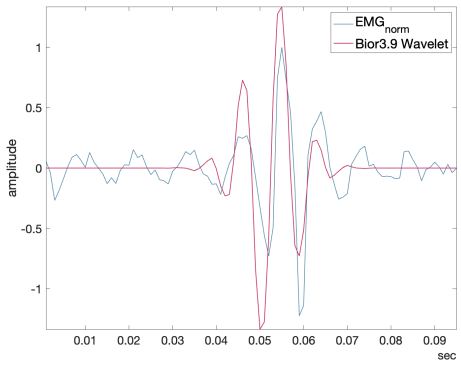

Fig. 1. Phasic activity observed in the mentalis muscle (blue line), and the bior3.9 mother wavelet selected for analysis and synthesis (red line).

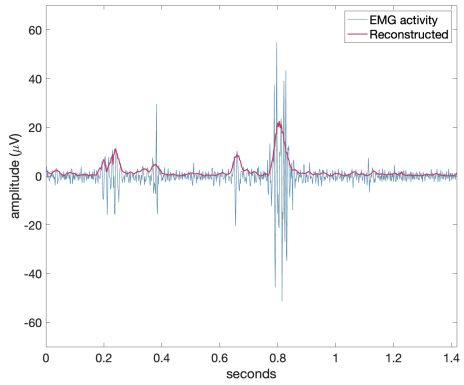

Fig. 2. Wavelet-based synthesis of a portion of EMG signal including phasic activity; the reconstructed signal is shown as solid red line.

probabilistic consensus of the four scorers was obtained for each 3-s mini-epoch and employed in this study as ground-truth for the final score comparison.

III. METHODS

A. Signal Pre-Processing

This study included only EMG recordings from the mentalis and bilateral FDS, as these are the muscles recommended by international guidelines for RWA scoring [8]. The EMG recordings were pre-processed with a Butterworth bandpass filter (50–300 Hz) and a notch filter for powerline rejection (50 Hz). Additionally, EMG signals from the mentalis muscle were bandpass filtered in the range (5–10 Hz), to better characterize slow drifts due to breathing or snoring. Finally, resampling to 600 Hz was performed (finite-impulse response, antialiasing, low-pass filter) to ease the subsequent processing steps.

B. Feature Extraction

Feature extraction was carried out on the EMG recordings to morphologically characterize: (i) artefacts from elevated background tone (*Artefact-Bkg*), and (ii) artefacts from phasic activity (*Artefact-Phasic*), following two different analysis pipelines. For the sake of clarity, these also represent the two classification tasks explored in Section III-D.

The procedures are described in detail in the following paragraphs.

1) Artefacts vs Background Activity: For this task, features were extracted only from the EMG recorded at the mentalis muscle, as elevated background tone is rarely seen in the FDS muscles [9]. Features were derived in the temporal (TD), frequency (FD), and non-linear (NLD) domains. Feature extraction was conducted on 1-second mini-epochs to ensure stationarity; then, a set of statistics (mean, median, mode, 25^{th} and 75^{th} percentiles, interquartile range (IQR) and interdecile range (IDR)) was obtained from each 3-second epoch and employed as independent features in the classification step. For the sake of clarity, the selected mini-epochs included on average 26.2 $\pm$ 1.1 % artefacts on a 4-scorer consensus. Table I reports the set of extracted features.

2) Artefact vs Phasic Activity: This step first envisaged the morphological characterization of phasic activity, followed by feature extraction. The procedures are detailed in the following paragraphs.

a) Morphological Characterization: As highlighted in [7], phasic activity is the most affected by artefacts, inevitably leading to interrater variability in RWA scoring. The selected mini-epochs featured 18.97 $\pm$ 6.37 % artefacts across all subjects (on four scorers consensus). Hence, this study first focused on accurately characterizing the morphology of phasic activity, to refine the feature extraction process. A Continuous Wavelet Transform (CWT) was applied to match the morphology of phasic activations, following an approach similar to [17], [18]. A biorthogonal mother wavelet (*bior3.9*) was selected for its similarity to the morphology of phasic activity (Figure 1). Then, 5-level decomposition was performed on the mini-epochs to obtain the approximation (cA) and detail (cD) coefficients. The Kernel Density Estimation (KDE) analysis highlighted the 75–150 Hz band (detail level 3) as most suitable for distinguishing artefacts from phasic activity. Finally, signal reconstruction at this

TABLE I

FEATURES EXTRACTED FOR THE MORPHOLOGICAL CHARACTERIZATION OF EMG SIGNALS IN THE ARTEFACT-BKG TASK. ◇: ADAPTED FROM CITED STUDY, †: FIRST PROPOSED IN THIS STUDY.

Feature	**Reference**
Time Domain	
Amplitude measures: mean, STD, skewness, kurtosis, range, maximum and minimum value, RMS	◇[10]
Zero Crossing Rate (ZCR)	◇[11]
Hjorth Parameters	◇[12]
Amplitude Percentiles (5^{th}, 10^{th}, 25^{th}, 75^{th}, 90^{th})	◇[10]
Form, Crest, and Impact Factors	◇[13]
Event Duration	†
Frequency and Non-Linear Domain	
Mean and median frequencies (MNF, MEDF), power	◇[14]
Spectral Edge Frequencies (SEF25, SEF75, SEF95)	◇[15], [14]
Entropy measures	◇[15], [16]

level was conducted, providing good fit to phasic activations (Figure 2). The reconstructed signal (mentalis and bilateral FDS muscles) was employed for the feature extraction step.

3) Feature Engineering: Two features were proposed to differentiate artefacts from phasic activity, extracted from 0.1-second epochs to match the shortest event duration for phasic activations. Given the strong morphological resemblance of the reconstructed signal with respect to phasic segments (Figure 2), a correlation-based metric was proposed. Namely, the Correlation-Based Index (CI), defined as the ratio between the cross-correlation of the original signal $x(t)$ and its reconstruction $y(t)$, and the signal's auto-correlation (Equation 1).

$$CI = \frac{R_{xy}}{R_{xx}} \tag{1}$$

The CI ranges from 0 to 1, with values approaching 1 indicating higher similarity. Additionally, to prevent bias due to low-amplitude portions of the EMG and accurately characterize the regions of interest, the 90$^{\text{th}}$ Energy Threshold (ET90) was introduced, since high-amplitude activity mostly lies in the top 10% of the signal amplitude. This is an energy-based metric, computed as the 90$^{\text{th}}$ percentile of the signal's integral, virtually representing a high-pass filter for the signal energy.

C. Feature Analysis and Selection

Statistical tests were conducted prior to the classification step to ensure the reliability of the proposed framework. Feature normality was assessed through the Shapiro-Wilk test. Then the Student's t test and the non-parametric Mann-Whitney U test were conducted on normally- and non-normally distributed features, respectively. Finally, the ReliefF feature selection method was adopted in each training fold [19] for the classification task (i) *Artefact-Bkg*. As the classification task (ii) *Artefact-Phasic* only envisaged two features (CI and ET90), no feature selection was performed.

D. Classification

Supervised Machine Learning (ML) models were employed for the two classification tasks (*Artefact-Bkg* and *Artefact-Phasic*). Five classifiers were explored in task (i): Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbors (KNN), Random Forest (RF), and Linear Discriminant Analysis (LDA). For the classification task (ii) the SVM, KNN, RF, LDA were employed, along with an ensemble method (AdaBoost). Given the limited size of the dataset, a Leave-One-Subject-Out cross-validation (LOSO-CV) approach was adopted. Hyperparameters optimization (Grid Search approach) was performed to enhance model performance and allow for better generalization capability, with the F1 score used as metric for model comparison. Classification performance was assessed through Accuracy, Recall, Specificity, Precision, and F1 score.

IV. RESULTS

A. Statistical Analysis and Feature Selection

The Shapiro-Wilk test in task (i) revealed a non-normal distribution ($p < 0.001$), while features in task (ii) were normally distributed ($p > 0.05$). For the *Activity-Bkg* task, several features emerged as statistical significant (Mann-Whitney U test); Table II reports the top-10 ranked features selected by the ReliefF algorithm, across all tested folds. As observable, SEF25$_{mean}$ emerged as the most important feature for the discrimination.

TABLE II
FEATURES EMPLOYED IN THE *Artefact-Bkg* CLASSIFICATION, WITH STATISTICAL SIGNIFICANCE (MANN-WHITNEY U TEST).
$^{*}: p < 0.05$, $^{**}: p < 0.005$, $^{***}: p < 0.001$.

Feature	Source	p
SEF25$_{mean}$	Mentalis (50–300 Hz)	$< 0.001^{***}$
Kurtosis$_{mean}$	Mentalis (50–300 Hz)	$< 0.001^{***}$
95pctl$_{mean}$	Mentalis (50–300 Hz)	$< 0.001^{***}$
75pctl$_{mean}$	Mentalis (5–10 Hz)	0.005^{*}
5pctl$_{mean}$	Mentalis (50–300 Hz)	$< 0.001^{***}$
90pctl$_{mean}$	Mentalis (50–300 Hz)	$< 0.001^{***}$
MEDF$_{mean}$	Mentalis (50–300 Hz)	$< 0.001^{***}$
AbsPower$_{mean}$	Mentalis (5–10 Hz)	$< 0.005^{**}$
MEDF$_{mean}$	Mentalis (5–10 Hz)	$< 0.001^{***}$
Entropy$_{mean}$	Mentalis (50–300 Hz)	$< 0.001^{***}$

TABLE III
CLASSIFICATION TASK: ARTEFACT VS BACKGROUND ACTIVITY. PERFORMANCE METRICS OF THE CLASSIFIERS (%).

	DT	SVM	KNN	RF	LDA
Acc.	71.05±2.28	79.7±3.2	73.7±2.77	76.9±2.1	68.9±2.37
Rec.	71.3±2.9	80.5±2.9	74.7±2.3	78.9±2.35	66.1±3.6
Spec.	70.81±3.09	78.7±2.4	72.9±1.08	75.30±2.54	70.9±3.13
Prec.	68.6±2.6	76.9±2.0	71.1±1.87	74.05±1.45	66.5±2.98
F1	70.9±4.1	79.2±4.1	73.9±3.86	76.56±2.0	66.3±4.21

In classification task (ii) the CI emerged as a promising detection tool, with phasic activity segments featuring a value of 0.71 $\pm$ 0.05, and 0.42 $\pm$ 0.12 for artefacts, with a significantly different data distribution in the two classes ($p < 0.05$ in Student's t-test).

B. Classification Performance

Table III reports the classification performance for the *Activity-Bkg* classification task. The SVM classifier emerged as the best model, achieving a 79.7% and 79.2% Accuracy and F1 score, respectively. Overall, the explored classifiers attained an average validation Accuracy of 74.05 $\pm$ 4.3 %, suggesting good predictive power of the selected features in detecting artefacts from background activity.

The second classification task (*Artefact-Phasic*) reached good-to-excellent validation scores; Table IV displays the results. The best performance was obtained by the LDA classifier, featuring an F1 score of 86.3 $\pm$ 3.83 %, Accuracy of 85.5 $\pm$ 3.3 %, and the highest Recall of all models, with a value of 88.4 $\pm$ 1.9 %. In summary, the models reached an average Accuracy of 83 $\pm$ 1.99 % and F1 score of 84.1 $\pm$ 2.09 %, demonstrating the strong discriminative power of the two proposed novel predictors in the identification of artefacts from phasic activity.

TABLE IV

CLASSIFICATION TASK: ARTEFACT VS PHASIC ACTIVITY. PERFORMANCE METRICS OF THE CLASSIFIERS (%).

	SVM	**KNN**	**NB**	**LDA**	**AdaBoost**
Acc.	84.4±1.2	81.6±2.5	80.6±2.11	85.5±3.3	82.8±1.97
Rec.	88±1.8	84.5±2.0	83.8±2.14	88.4±1.9	86.7±3.06
Spec.	81.6±2.0	79.8±2.4	78.3±3.03	82.3±1.7	77.9±2.9
Prec.	80.9±2.1	78.1±1.6	78.2±4.25	84.4±1.5	81.5±1.6
F1	84.5±1.4	85.1±2.3	80.7±1.9	86.3±3.83	83.9±2.61

V. DISCUSSION AND CONCLUSIONS

This study proposed a framework for the automatic detection of artefacts in EMG recordings in REM sleep through supervised ML models based on objective descriptors of EMG morphology and a matched-wavelet approach. Although based on a simple analysis pipeline, the framework effectively discriminated artefacts from elevated background and phasic activity, with average Accuracies of 74% and 83%, respectively, across the explored models. Furthermore, the good-to-excellent performance obtained in the *Artefact-Phasic* task reveals the high predictive power of the proposed novel metrics. The slightly lower performance observed in the *Artefact-Bkg* classification task might be due to the morphological similarity between actual activity and artefacts, particularly those caused by breathing or snoring, which are inherently challenging to detect; nevertheless, the proposed method demonstrates promising potential. It is worth noting that this is the first study tackling automatic artefact detection in REM sleep through simple and lightweight EMG metrics; therefore, a direct comparison with the previous literature is impracticable. However, this highlights the novelty of the proposed framework, and the performance obtained supports its potential for further refinements and clinical implementation. Future work should address the limitations of this study. First, the analysis should include a larger population to improve generalizability. Second, the framework, designed and validated on manually selected epochs, should be tested on additional data, possibly without manual pre-selection, to minimize the risk of overfitting. Furthermore, validation with clinical RWA metrics is needed to ensure robustness. Further investigations could explore hybrid signal decomposition techniques (e.g., variational mode decomposition [20]) for artefact detection; combining such approaches with morphological descriptors may contribute to more robust detection models. Despite preliminary, the results demonstrate the effectiveness of a low-computational-cost method for automated artefact detection in EMG. This offers a significant step toward fully automatic quantification of RWA, significantly reducing PSG scoring times and minimizing interrater variability, and paving the way for more efficient clinical assessments.

ACKNOWLEDGMENT

This study is part of the project NODES, funded by the Italian Ministry for Universities and Research – M4C2 1.5 of PNRR, under the EU NextGenerationEU program (Grant ECS00000036).

REFERENCES

[1] M. J. Sateia, "International classification of sleep disorders," *Chest*, vol. 146, no. 5, pp. 1387–1394, 2014.

[2] B. Högl and A. Stefani, "Rem sleep behavior disorder (rbd): Update on diagnosis and treatment," *Somnologie*, vol. 21, no. Suppl 1, p. 1, 2017.

[3] B. Frauscher *et al.*, "Quantification of electromyographic activity during rem sleep in multiple muscles in rem sleep behavior disorder," *Sleep*, vol. 31, no. 5, pp. 724–731, 2008.

[4] R. B. Berry *et al.*, "The aasm manual for the scoring of sleep and associated events," *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, vol. 176, no. 2012, p. 7, 2012.

[5] M. Cesari and I. Rechichi, "Automatic and machine learning methods for detection and characterization of rem sleep behavior disorder," in *Handbook of AI and Data Sciences for Sleep Disorders*, pp. 197–217, Springer, 2024.

[6] B. Frauscher *et al.*, "Validation of an integrated software for the detection of rapid eye movement sleep behavior disorder," *Sleep*, vol. 37, no. 10, pp. 1663–1671, 2014.

[7] M. Cesari *et al.*, "Flexor digitorum superficialis muscular activity is more reliable than mentalis muscular activity for rapid eye movement sleep without atonia quantification: A study of interrater reliability for artifact correction in the context of semiautomated scoring of rapid eye movement sleep without atonia," *Sleep*, vol. 44, no. 9, p. zsab094, 2021.

[8] M. Cesari *et al.*, "Video-polysomnography procedures for diagnosis of rapid eye movement sleep behavior disorder (rbd) and the identification of its prodromal stages: guidelines from the international rbd study group," *Sleep*, vol. 45, no. 3, p. zsab257, 2022.

[9] B. Frauscher, A. Iranzo, *et al.*, "Normative emg values during rem sleep for the diagnosis of rem sleep behavior disorder," *Sleep*, vol. 35, no. 6, pp. 835–847, 2012.

[10] N. Cooray *et al.*, "Detection of rem sleep behaviour disorder by automated polysomnography analysis," *Clinical Neurophysiology*, vol. 130, no. 4, pp. 505–514, 2019.

[11] K. Šušmáková and A. Krakovská, "Discrimination ability of individual measures used in sleep stages classification," *Artificial intelligence in medicine*, vol. 44, no. 3, pp. 261–277, 2008.

[12] S. Motamedi-Fakhr, M. Moshrefi-Torbati, M. Hill, C. M. Hill, and P. R. White, "Signal processing techniques applied to human sleep eeg signals—a review," *Biomedical Signal Processing and Control*, vol. 10, pp. 21–33, 2014.

[13] I. Rechichi, F. Amato, A. Cicolin, and G. Olmo, "Single-channel eeg detection of rem sleep behaviour disorder: The influence of rem and slow wave sleep," in *International Work-Conference on Bioinformatics and Biomedical Engineering*, pp. 381–394, Springer, 2022.

[14] I. Rechichi, A. Iadarola, M. Zibetti, A. Cicolin, and G. Olmo, "Assessing rem sleep behaviour disorder: From machine learning classification to the definition of a continuous dissociation index," *International Journal of Environmental Research and Public Health*, vol. 19, no. 1, p. 248, 2021.

[15] I. Rechichi, M. Zibetti, L. Borzì, G. Olmo, and L. Lopiano, "Single-channel eeg classification of sleep stages based on rem microstructure," *Healthcare Technology Letters*, vol. 8, no. 3, p. 58, 2021.

[16] U. R. Acharya, F. Molinari, S. V. Sree, S. Chattopadhyay, K.-H. Ng, and J. S. Suri, "Automated diagnosis of epileptic eeg using entropies," *Biomedical signal processing and control*, vol. 7, no. 4, pp. 401–408, 2012.

[17] M. Cesari, J. Mehlsen, A.-B. Mehlsen, and H. B. D. Sorensen, "A new wavelet-based ecg delineator for the evaluation of the ventricular innervation," *IEEE journal of translational engineering in health and medicine*, vol. 5, pp. 1–15, 2017.

[18] G. Olmo, F. Laterza, and L. L. Presti, "Matched wavelet approach in stretching analysis of electrically evoked surface emg signal," *Signal Processing*, vol. 80, no. 4, pp. 671–684, 2000.

[19] R. J. Urbanowicz *et al.*, "Relief-based feature selection: Introduction and review," *Journal of biomedical informatics*, vol. 85, pp. 189–203, 2018.

[20] D. Pachori *et al.*, "Detection of atrial fibrillation from ppg sensor data using variational mode decomposition," *IEEE Sensors Letters*, vol. 8, no. 3, pp. 1–4, 2024.