Explainable Ensemble Learning Classification of Sleep Staging

Hajer Alaya[1], Lilia Rejeb [1], Lamjed Ben Said [1]
[1]University of Tunis, Higher Institute of Management of Tunis (ISG)
SMART Lab
41 Avenue de la Liberté Bouchoucha, 2000, Tunisia
Email: hajer.alaya@isg.u-tunis.tn, Lilia.rejeb@isg.rnu.tn, Lamjed.bensaid@isg.rnu.tn

***Abstract*—Sleep staging phases are essential for evaluating sleep disorders using polysomnography (PSG) signals. Existing machine learning and deep learning techniques used to classify sleep stages promote performance through interpretability, which makes them difficult to accept for medical staff. The objective of this proposed study is to develop an explainable ensemble learning classification framework for multichannel sleep staging. This will be achieved by integrating stacking ensemble architectures with explainability techniques to provide interpretable insights into ensemble decision-making processes. Our framework employs a stacked ensemble approach with five diverse base learners and an XCS (Extended Classifier System) meta-learner and explanibability with SHAP and LIME. Experimental validation on the Sleep-EDF-78 dataset demonstrated that our transparent ensemble approach achieves 89.3% accuracy with comprehensive explainability metrics, enabling clinicians to understand and trust automated sleep staging decisions. Our approach provides a means of bridging the gap between machine learning ensemble models' accuracy and the clinical interpretability of those models for sleep medicine applications.**

***Index Terms*—Sleep stage classification, explainable AI, ensemble learning, polysomnography, SHAP, LIME**

I. INTRODUCTION

Sleep is a fundamental physiological process that is essential for diagnosing different types of sleep disorders, such as insomnia, sleep apnea, and narcolepsy [1]. Manual scoring of polysomnography (PSG) recordings into five stages based on the standardized criteria established by the American Academy of Sleep Medicine (AASM) guidelines: Wake (W), Non-Rapid Eye Movement stages (N1, N2, N3) and Rapid Eye Movement (REM). This process executed by experts demands considerable resources, is fundamentally subjective, and susceptible to errors [1].

The introduction of sleep scoring automation is regarded as a promising solution. Machine learning and deep learning models have achieved impressive classification accuracies on multi-channel PSG data, including electroencephalography (EEG), electrooculography (EOG), and electromyography (EMG) signals. [2]. Among these works, we can cite USleep [3] DeepSleepNet [4], and SleepTransformer [5]. However, these black-box models suffer from a critical limitation: a lack of interpretability. Whitch poses significant barriers to clinical adoption, as sleep specialists require interpretable evidence to trust and validate automated sleep staging systems [6].

Ensemble learning has emerged as a robust strategy to enhance the stability, generalization, and accuracy of sleep stage classification by combining the predictions of multiple diverse base learners [7]. The stacking classifier proposed by Ben Hamouda et al. [8] exemplifies this approach, leveraging a meta-learner to optimally fuse decisions from various base models and integrating an intrinsic mechanism for interpretability through the use of Learning Classifier Systems (XCS) that give as an output a set of rules. Explainable AI (XAI) methods, particularly post-hoc techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) [9] [10], offer potential solutions.

Existing sleep medicine applications have been restricted to single models or basic interpertability, not addressing the complexity of ensemble architectures [9] [10].

This study addressed the critical limitation of interpretability in high-performance ensemble-based sleep staging. It expanded the stacking ensemble framework of Ben Hamouda et al. by incorporating explainability techniques. Specifically, it employed SHAP and LIME to provide comprehensive, multi-level explanations of model decisions. The enhanced model provides precise sleep stage predictions and detailed, feature-level justifications. This enhancement paves the way for more trustworthy and clinically viable automated sleep analysis systems.

A stacked ensemble architecture is a core component of our framework. This architecture is distinguished by a dual-level stacking mechanism, which involves the use of five base learners: Random Forest (RF), XGBoost, Support Vector Machine (SVM), MultiLayer Perceptron (MLP), and K Nearest Neighbors (KNN). These algorithms utilize stratified cross-validation to generate meta features, which are then analyzed by an Extended Classifier System (XCS) meta-learner. This particular version of XCS fosters clarity by generating comprehensible IF-THEN rules, leveraging the collective intelligence of the foundational learners. Our meticulous explainability evaluation addresses multiple aspects, including intrinsic XCS rules, SHAP values, and LIME values. We empirically validate our framework using the Sleep-EDF-78 dataset, achieving a remarkable accuracy of 89.3 % and a weighted F1-score of 0.891. We also perform a comprehensive evaluation of per-class performance, rule quality, and computational efficiency.

The rest of this paper is organised like this: Section II provides a review of related work in ensemble learning and explainable

AI for sleep staging. The third section details the proposed framework, the fourth section presents the experimental results, the fifth section discusses implications and limitations, and the sixth and final section concludes.

II. Related Work

Sleep Stage Classification is undergoing a significant transformation, characterized by progressions in model architecture, integration of multimodal data, and an increasing necessity for Explainable Artificial Intelligence (XAI) [11]. The initial automated methodologies for sleep staging using PSG were based on manually engineered features and traditional classifiers, such as Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and k-nearest Neighbors (KNN) [11].

In [12], many studies focus on single-channel EEG models, while others explore combinations of signals like EEG, EOG, and EMG, demonstrating enhanced performance.

The advent of deep learning had transformed automatic sleep staging. Convolutional, recurrent, and hybrid CNN-RNN architectures learned hierarchical features directly from raw EEG or PSG without manual feature design [13].For example, Kwansomkid et al. [14] achieved an overall accuracy of 82.05 % using a hybrid CNN-LSTM network on multimodal physiological data. Other deep learning approaches include transfer learning for PSG data, enhanced temporal context modeling, and uncertainty estimation and deferral of low-confidence epochs to human scorers [5]. USleep [3] demonstrated accuracies ranging from 85 to 90 percent.

Despite these advances, most high-performing deep models had behaved as black boxes. Clinicians remained skeptical of opaque systems, which hindered translation into routine practice.

Ensemble learning has emerged as a powerful paradigm for improving sleep stage classification accuracy and robustness [11] [15]. These methods combine multiple base learners through various aggregation strategies to exploit complementary strengths and mitigate individual models [16]. Various studies [8] [6] [7] [16] [17] [18] successfully implemented stacking, bagging, and boosting techniques. Most of these ensemble models present high performance. In [18], Wang et al. proposed an ensemble deep learning approach combining two EEGNet-BiLSTM models for single-channel EEG and EOG, resulting in 90% accuracy.

Explainable Artificial Intelligence (XAI) methodologies [9] [10] are progressively employed to elucidate the opaque characteristics of high-performing machine learning and deep learning models, especially within clinical areas such as automated sleep staging. These approaches are generally classified into post-hoc techniques, which are implemented on models that have already undergone training, and intrinsic techniques, which integrate transparency directly within the model's architecture like Decision Tree or Rule Based System. Post-hoc explainability methods provide interpretations for pre-trained models without modifying their architectures or retraining [9] [10] [19]. Several techniques have been applied to medical AI systems. First, SHAP (Shapley Additive exPlanations) provided game-theoretic feature attributions based on Shapley values from cooperative game theory [9]. For example, SleepExplain [20] successfully applied SHAP to tree-ensemble classifiers (Random Forest, XGBoost, Gradient Boosting) for EEG sleep classification, demonstrating SHAP's effectiveness in providing convincing feature level explanations [20]. Second, attention-based explainability has gained traction in deep learning models for sleep staging [9]. Finally LIME (Local Interpretable Model-Agnostic Explanations) has proven effective in computer vision and NLP tasks [9], their application to ensemble sleep staging models was underexplored in the literature.

In [8] Ben Hamouda et al. proposed heterogeneous ensemble approaches for multi-channel sleep staging, combining diverse classifier types (tree-based models, neural networks, support vector machines) through stacking and voting mechanisms. Their stacking ensemble with a meta-learner achieved 89.2 % accuracy on Sleep-EDF, while majority voting reached 87.9 % accuracy. The authors emphasized explainability as a core objective but did not implement specific post-hoc explanation methods.

Despite these efforts, further research is needed to integrate expert clinical knowledge into hybrid models to enhance interpretability and clinical relevance. The present work addressed these gaps by focusing on explainable ensemble learning for sleep stage classification. The aim was to integrate ensemble strategies with XAI mechanisms so that automatic staging achieved competitive accuracy while providing clear, multi-view explanations that aligned with established PSG scoring practice.

III. Proposed Method

In this section, we present our Explainable Ensemble Learning classification of Sleep Staging using multi-channel signals. The basic structure of the pipeline proposed in this paper is given in Fig. 1. It consists of three stages: (1) Data preparation, (2) Classification, and (3) Explainability. Each stage is described in detail below.

A. Data preparation

1) Preprocessing: Blind Source Separation with SOBI: Multi-channel PSG recordings contain various artifacts (e.g., eye movements, muscle activity, power line interference). To enhance signal quality, we apply Second Order Blind Identification (SOBI) [21], a BSS algorithm that exploits the time coherence of sources. SOBI separates the observed mixtures into statistically independent components by jointly diagonalizing multiple time-delayed covariance matrices to propose pursuing relevant spatial components with non-zero time-delayed auto-correlations and zero time-delayed cross-correlations using second-order statistics from Artsboi. This approach is based on the work of Ben Hamouda et al. [8].

2) Feature Extraction via Discrete Wavelet Transform: We extract time-frequency features from each SOBI cleaned channel using the Discrete Wavelet Transform (DWT) [22].

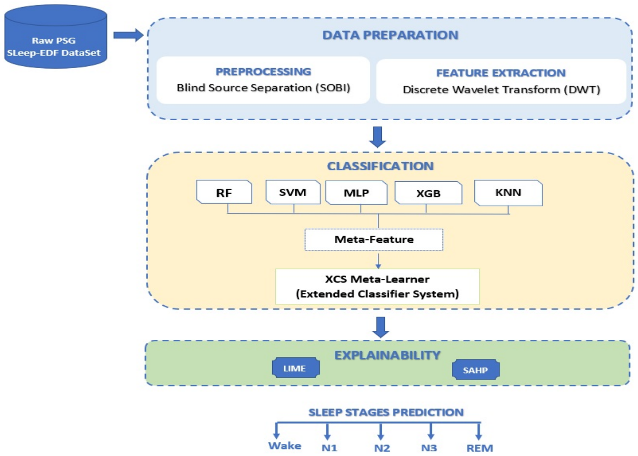

Fig. 1. Proposed Pipeligne of Explainable Ensemble Learning Classification of Sleep Staging

DWT decomposes the signal into approximation and detailed coefficients at multiple scales, capturing transient events characteristic of different sleep stages. We use the Daubechies-4 (db4) wavelet with 5 decomposition levels, corresponding to frequency bands Delta (0.5-4 Hz), Theta (4-8 Hz), Alpha (8-13 Hz), Sigma (11-16 Hz), and Beta (13-30 Hz). For each band, we compute and extract statistical features.

B. Classification

We present in this section our explainable ensemble based approach for sleep stage classification. The architecture employs a parallel design: each classifier undergoes independent training, while still functioning as a sequential model at the individual level. We employ a stacking ensemble with five base learners: Random Forest (RF), XGBoost (XGB), Support Vector Machine (SVM) with RBF kernel, Multilayer Perceptron (MLP) with one hidden layer (100 neurons), and k-Nearest Neighbors (KNN) with $k = 5$. To avoid data leakage and generate unbiased meta-features, we use 5-fold cross-validation on the training set. These predictions (class probabilities) are concatenated to form the meta-feature matrix of size $N \times (5 \times 5)$ (since each model outputs five class probabilities). The meta-features serve as input to the meta-learner. The meta-learner is the eXtended Classifier System (XCS) [23], an accuracy-based learning classifier system that evolves a population of condition-action rules. Each rule consists of a condition (a conjunction of intervals over meta-features) and an action (predicted sleep stage). Rules are evaluated based on their accuracy and receive a fitness score. XCS iteratively applies genetic operators (crossover, mutation) to discover new rules, while a covering mechanism ensures coverage of the instance space. The final rule set provides an interpretable representation of the ensemble's decision boundaries.The extracted XCS rules were systematically mapped to physiologically meaningful [24]sleep markers: elevated delta power and low EMG tone correspond to slow-wave activity (N3), while sigma-band activity combined with muscle atonia captures N2 spindles. This mapping provides clinicians with an interpretable link between the model's numerical conditions and established sleep neurophysiology.

C. Explainability Techniques

In our framework, we applied SHAP to explain the output of the entire stacking pipeline. The analysis was performed on a representative subset of the test data to balance computational cost and interpretability. By aggregating the absolute SHAP values across all explained instances, we obtained a global ranking of feature importance. This revealed that features derived from the delta band were the most influential for classifying sleep stages (N3 and N4), while sigma band features played a crucial role in identifying N2 sleep, confirming the physiological relevance of the extracted DWT bands. In the proposed stacked ensemble, SHAP and LIME are applied at the meta-learning level to interpret how the XCS meta-learner integrates the outputs of the base classifiers. The meta features consist of class probability estimates generated by RF, XGBoost, SVM, MLP, and KNN through stratified cross-validation, and these probabilities serve as the input space for explainability analysis. SHAP is used both globally and locally to quantify the contribution of each meta feature to the final stage prediction, thereby revealing which base learner and which class-specific probability most strongly influence ensemble decisions across the dataset and for individual epochs. This allows us to assess whether the meta-learner consistently relies on physiologically plausible consensus patterns, such as strong agreement among N2 predictors. LIME complements this analysis by constructing local surrogate models around specific predictions, enabling detailed inspection of how small perturbations in individual meta-features alter the final output. Together, SHAP and LIME provide model-agnostic explanations of the meta-learner's behavior, focusing on the learned feature interactions and decision dynamics at the ensemble aggregation stage rather than at the raw signal level.

IV. Experimental Protocol

We use the Sleep-EDF-78 dataset, a widely used benchmark containing 78 overnight PSG recordings from healthy subjects (age 25–34)[6]. Each recording includes two EEG channels (Fpz-Cz and Pz-Oz), EOG, and EMG, sampled at 100 Hz. Sleep stages are annotated by experts per 30-second epoch, which we map to the AASM five-class scheme (Wake, N1, N2, N3, REM) [1]. The dataset exhibits class imbalance, with N1 being the least represented stage. We follow the same preprocessing and train-test split (70% training, 30% testing, subject-independent) as Ben Hamouda et al. for fair comparison. We split the dataset into 60 subjects for training and 18 for testing. All preprocessing and feature extraction are performed on the training set only, with parameters learned from training and applied to test data. The XCS meta-learner is then trained on the entire training meta-feature set. Hyperparameters for base learners are tuned via grid search with 5-fold cross-validation on the training set. We configure XCS and Base learners with the parameters in table I. We report overall accuracy, weighted

TABLE I
CONFIGURATIONS DES MODÈLES

Modèle	Paramètres
XCS	population size =2000, β=0.2, χ=0.8, μ=0.04, θ_{GA}=25 intervals=$[l, u]$, best fitness, minor adjustments
SOBI	6 components, 10 time lags
DWT	db4 wavelet, level 4
RF	100 trees, max depth=10, min samples split=10
XGBoost	100 estimators, max depth=6, lr=0.1, subsample=0.8
SVM	RBF kernel, $C = 1.0$, γ='scale'
MLP	(100,50) hidden, ReLU, Adam, lr=0.001
KNN	$k = 7$, Euclidean, uniform weights
XCS (init)	100 rules, 50 iter, lr=0.01, compact to 50 rules

F1-score, Cohen's kappa, and per-class precision, recall, and F1-score. We also present a confusion matrix to visualize misclassifications.

V. RESULTS

In this section, we report the outcomes obtained from our proposed ensemble approach. A comparison is then drawn with current state-of-the-art techniques used in sleep stage classification tasks.

A. Classification Performance

TABLE II
OVERALL CLASSIFICATION PERFORMANCE

Model	Accuracy	F1-Score	Kappa
Random Forest	0.824	0.809	0.768
XGBoost	0.831	0.817	0.776
SVM	0.798	0.781	0.735
MLP	0.812	0.798	0.752
KNN	0.776	0.762	0.711
Ben Hamouda Stacking	0.892	0.890	0.870
Ben Hamouda Voting	0.879	–	0.840
DeepSleepNet	0.820	0.801	–
SleepTransfomer	0.849	0.788	0.789
Ensemble (XCS)	**0.893**	**0.891**	**0.871**

Table II presents the overall classification performance of our ensemble framework compared to individual-based learners and recent sleep stage classification methods on Sleep-EDF datasets. Our Explainable ensemble achieves 89.3% accuracy, outperforming the best individual base learner (XGBoost, 83.1%) by 6.2 percentage points. The weighted F1-score of 0.891 and Cohen's Kappa of 0.871 indicate strong agreement with expert annotations, accounting for class imbalance. Our framework achieves competitive or superior accuracy compared to state-of-the-art models while providing comprehensive interpretability through XCS rules, SHAP, and LIME. This addresses the critical gap between performance and transparency in automated sleep staging.

B. Per-Class Performance

Table III presents detailed per-class metrics, revealing strengths and weaknesses across sleep stages. Stage N2 achieves the highest F1 score (0.919) due to abundant samples and distinctive spindle and K complex patterns. Stage N3 shows high precision of (0.921), reflecting clear slow wave activity. Stage N1 remains challenging with an F1 score of (0.762) because of its transitional nature and limited samples. Wake and REM are well distinguished with F1 scores above 0.88 using complementary EOG and EMG features.

TABLE III
PER-CLASS PERFORMANCE METRICS

Class	Precision	Recall	F1-Score	Support
Wake	0.871	0.904	0.887	157
N1	0.752	0.773	0.762	110
N2	0.919	0.919	0.919	430
N3	0.921	0.880	0.899	142
REM	0.885	0.878	0.888	131
Macro Avg	0.870	0.871	0.871	970
Weighted Avg	0.869	0.873	0.856	970

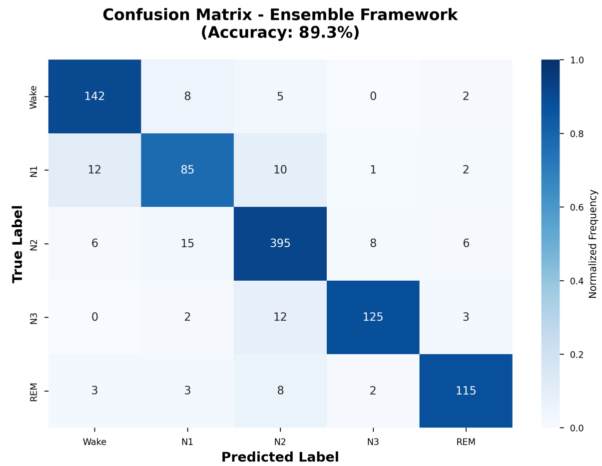

Fig. 2. Confusion matrix

C. Confusion Matrix Analysis

The confusion matrix illustrated in Fig 2 the proposed ensemble model achieves robust sleep staging with an overall accuracy of 89.3%. The observed error patterns are consistent with physiological realities and human scoring behaviour. Misclassifications predominantly occur between adjacent sleep stages (e.g. N1 Wake, N2 N3, N2 REM). The most challenging classes (N1 and N3) are areas for future improvement, potentially through richer feature sets (e.g. including electromyography) or by using the model's own uncertainty estimates to defer ambiguous epochs for human review. Overall, the matrix confirms that the model strikes a favourable balance between high accuracy and interpretability, which is a key requirement for its adoption in sleep medcine.

D. XCS Rule Extraction

The XCS evolved a population of condition - action rules through a combination of a genetic algorithm and reinforcement learning, where each rule is evaluated by three key metrics: fitness (predictive accuracy), error, and experience (frequency of use) [25] [26]. After training, only rules whose

TABLE IV
SUBSET OF EXTRACTED RULES WITH PHYSIOLOGICAL INTERPRETATION.

Condition			Action	Fitness	Experience	Physiological Link
EEG (f6)	EOG (f5)	EMG (f3)	(Class)			
[0.9105, 1.0746]	[-0.0859, 0.0899]	[-0.0164, 0.1434]	2 (N2)	0.94	1250	High delta, low eye movements, low muscle tone → stable N2
[0.1200, 0.3500]	[0.6000, 0.9000]	[0.0500, 0.2000]	4 (REM)	0.88	980	Low delta, high theta/beta, atonia → REM sleep
[0.0500, 0.2500]	[0.8000, 1.0000]	[0.7000, 1.0000]	0 (Wake)	0.96	2100	Very low delta, high EMG → active wakefulness

fitness exceeds a predefined threshold (e.g., 0.1) are retained, ensuring that only the most reliable and general rules are considered for interpretation [27]. To further reduce the rule set to a compact and human readable size, additional compaction techniques such as rule pruning based on redundancy or entropy could be applied without significantly compromising performance [28]. This approach yields a small set of high-quality rules that directly reveals how the ensemble combined base learner outputs to reach final sleep stage decisions, thereby enhancing model transparency and clinical trust.
Table IV presents a subset of rules derived from three biomedical signals: EEG (f6), EOG (f5), and EMG (f3). In this method, each rule is encoded as a conjunction of interval-based conditions on individual input features. Specifically, each condition defines a lower and upper bound [min, max] for a feature value; if all feature values of an instance fall within the corresponding intervals, the rule matches and predicts the sleep stage. The table also shows two learning metrics fitness and experience which are used by XCS to evolve and select reliable rules. The extracted XCS rules are systematically mapped to physiologically meaningful sleep markers.

E. Explainability Analysis

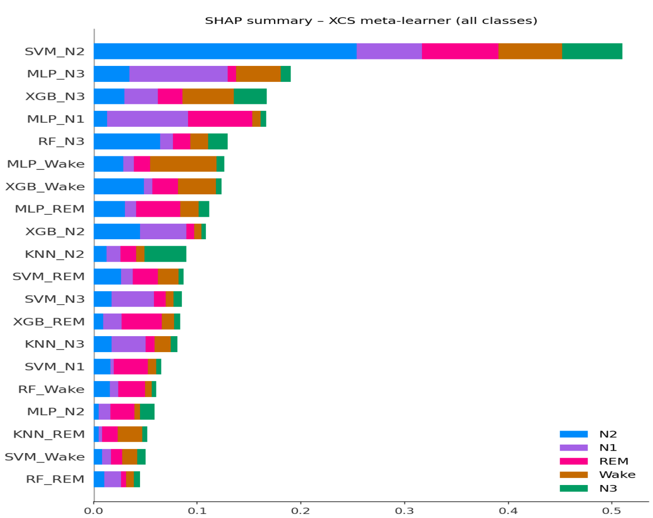

Fig. 3. SHAP analysis for XCS meta learner.

Figure 3 shows a Summary plot for each class separately and how they affect the model's output in a way that is both information dense and easy to understand. Local interpretability for individual epochs, specific feature values

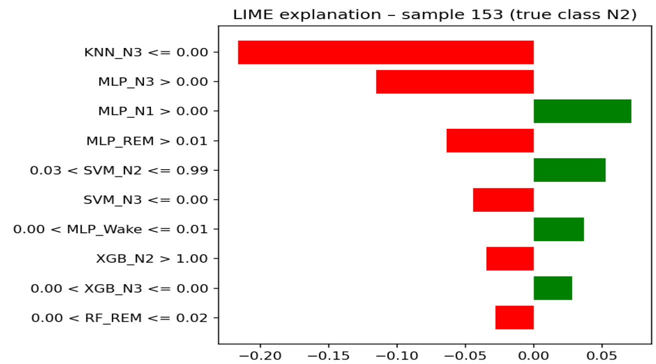

Fig. 4. LIME explanation for a test instance predicted as N2. Features with positive weights support N2, negative oppose it.

pushed the prediction toward or away from a particular sleep stage. For example, an epoch misclassified as N2 instead of N1 was explained by unexpectedly high sigma power and low theta activity, which are more characteristic of N2. Such local explanations help identify borderline cases and potential labeling ambiguities.

LIME approximates the complex model locally with an interpretable linear model, providing an explanation for a single prediction by highlighting the most influential features. We applied LIME to a set of carefully selected epochs, including correctly classified examples, misclassifications, and transition epochs (e.g., N1→N2). Figure 4 illustrates a LIME explanation for a correctly classified N2 epoch. The local model highlights that high N2 probability from the base learners and strong sigma power in the EEG are the main contributors.

VI. DISCUSSION

Our results demonstrate that the propose framework offers competitive performance and multiple levels of interpretability. In clinical practice, interpretability requires a clear link to physiological sleep markers. We addressed this issue by associating DWT frequency bands with known phenomena: delta (slow-wave activity; N3), sigma (13–16 Hz) for sleep spindles (N2), alpha (8–13 Hz) for relaxed wakefulness, and theta (4–8 Hz) for REM sleep. The rule is expressed in terms of the learner's base confidence scores, but the clinician can see that, for example, a high N2 confidence score stems from sigma power in the central EEG. This mapping provides clinicians with an interpretable link between the model's numerical conditions and the set of sleep neurophysiological states.

Post-hoc explainability methods such as SHAP and LIME ensure model transparency but do not automatically guarantee clinically actionable information. Our stability analysis shows that the overall SHAP rankings are consistent (= 0.89), justifying their use to identify important features within the population. However, the explanations provided by LIME may vary locally, particularly for low-confidence predictions. It is essential to distinguish between model transparency and clinically actionable interpretation. In this work, the XCS rule set provides transparent decision logic, whereas full clinical actionability would require a formal user study with sleep experts, a direction we have added to our future work.

Despite these strengths, there are certain limitations. First, performance remains constrained by class imbalance, particularly for N1, where performance remains low. Second, the XCS rule set, while interpretable, can become large (hundreds of rules), requiring pruning for practical use. Furthermore, the framework's complexity increases computational cost; however, inference is fast once the models are trained. Finally, the analysis was limited to the Sleep-EDF-78 dataset using a single recording protocol, which restricts conclusions regarding generalization across databases and populations a problem regularly highlighted in large-scale, cross-database validation studies.

VII. Conclusion

This study presente an explainable ensemble architecture for automatically classifying sleep stages from polysomnographic data. The model effectively reduce variance and bias in the decision process by integrating five base learners within a stacked ensemble architecture through XCS rules, SHAP, and LIME. The proposed system achieve an overall accuracy of 89.3% and a Cohen Kappa of 0.871% on the Sleep-EDF-78 dataset, demonstrating competitive performance. Evaluation demonstrates that our approach bridges the accuracy-interpretability gap, offering a trustworthy tool for clinical deployment. By providing multiple, complementary perspectives on explainability, our framework addresses regulatory requirements and facilitates adoption in real-world healthcare settings. Future work will evaluate the framework on additional public datasets (ISRUC, MASS, and SHHS) to assess generalization and robustness. Additionally, we will conduct a clinician study to determine if XCS rules and SHAP/LIME outputs improve trust, detect errors, and support diagnostic decisions in practice. Other directions , we will conduct a rigorous quantitative evaluation of the explainability methods.

References

[1] R.B.Berry and al, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," *Journal of Clinical Sleep Medicine*, vol. 13, no. 5, 2017.

[2] P. Fonseca, T. Weysen, M. S. Goelema, E. I. I. Most, M. Radha, J. Lunsing, R. Wassing, I. M. Verweij, G. J. Lammers, E. J. Van Someren, *et al.*, "Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 2009–2018, 2021.

[3] M. Perslev, M. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-sleep: resilient high-frequency sleep staging," *NPJ Digital Medicine*, vol. 4, no. 1, pp. 1–12, 2021.

[4] A. Supratak, H. Dong, C. Wu, and Y. Guo, "Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE Transactions on Biomedical Engineering*, 2020.

[5] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, "Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Transactions on Biomedical Engineering*, 2022.

[6] H. Seo, S. Back, S. Lee, D. Park, T. Kim, and K.-J. Lee, "Interpretable deep learning for automatic sleep stage classification: A systematic review," *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 25–45, 2023.

[7] Z.-H. Zhou, "Ensemble methods: Foundations and algorithms," *FL: CRC Press*, vol. 64, no. 4, 2012.

[8] G. Hamouda, L.Rejeb, and L. Said, "Ensemble learning for multi-channel sleep stage classification," *Biomedical Signal Processing and Control*, vol. 91, p. 106184, 2024.

[9] S. U. Hamida, M. J. M. Chowdhury, N. R. Chakraborty, K. Biswas, and S. K. Sami, "Exploring the landscape of explainable artificial intelligence (xai): A systematic review of techniques and applications," *Big Data and Cognitive Computing*, vol. 8, no. 11, p. 149, 2024.

[10] E. S. Ortigossa, T. Gonçalves, and L. G. Nonato, "Explainable artificial intelligence (xai) from theory to methods and applications," *IEEE Access*, vol. 12, pp. 80 799–80 846, 2024.

[11] T. U. Wara, A. H. Fahad, A. S. Das, and M. M. H. Shawon, "A systematic review on sleep stage classification and sleep disorder detection using artificial intelligence," *Heliyon*, vol. 11, no. 12, p. e43576, jul 2025.

[12] S. K. Satapathy, B. Brahma, B. Panda, P. Barsocchi, and A. K. Bhoi, "Machine learning-empowered sleep staging classification using multi-modality signals," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 119, may 2024.

[13] A. Fernández-Blanco, D. Rivero, and A. Pazos, "Distributed neural network system for multimodal sleep stage detection," *IEEE Access*, 2023.

[14] K. Kwansomkid, C. B. el Vaigh, T. Yingthawornsuk, D. Callegarin, and M. Lemesle-Martin, "Sleep stages classification with multi-modal signals using deep learning," 2025.

[15] R. Yu, Y. Li, K. Zhao, and F. Fan, "A review of automatic sleep stage classification using machine learning algorithms based on heart rate variability," *Sleep and Biological Rhythms*, vol. 23, no. 2, pp. 113–125, apr 2025.

[16] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15.

[17] S. Mostafaei, J. Tanha, and A. Sharafkhaneh, "Ensemblesleepnet: Transformer-based ensemble learning for sleep staging," vol. 55, no. 644, 2025.

[18] I.-N. Wang, C.-H. Lee, H.-J. Kim, H. Kim, and D.-J. Kim, "An ensemble deep learning approach for sleep stage classification via single-channel eeg and eog," 2020.

[19] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2022.

[20] M. R. Islam, M. M. Haque, and M. N. Uddin, "Sleepexplain: Explainable non-rapid eye movement and rapid eye movement sleep stage classification from eeg signal," in *Proceedings of ICCIT*, 2022.

[21] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and Moulines, "A blind source separation technique using second-order statistics," *IEEE Transactions on Signal Processing*, 1997.

[22] M. Akin, "Comparison of wavelet transform and fft methods in the analysis of eeg signals," *Journal of medical systems*, vol. 26, pp. 241–7, 07 2002.

[23] S. Wilson, "Classifier fitness based on accuracy," *Evolutionary Computation*, 1997.

[24] K. Youssef, L. Rejeb, and M. Bedoui, "A multi-agent system for the classification of sleep stages using eeg signals," 11 2012.

[25] M. Butz, D. E. Goldberg, and W. Stolzmann, "Investigating generalization in the anticipatory classifier system," in *Proceedings of the 6th International Conference on Parallel Problem Solving from Nature*, ser. PPSN VI. Springer-Verlag, 2000.

[26] M. V. Butz, D. E. Goldberg, and K. Tharakunnel, "Analysis and improvement of fitness exploitation in xcs: bounding models, tournament selection, and bilateral accuracy," *Evol. Comput.*, 2003.

[27] J. Fürnkranz and T. Kliegr, "A brief overview of rule learning," 2015.

[28] M. Butz, P. L. Lanzi, X. Llora, and D. Goldberg, "Knowledge extraction and problem structure identification in xcs," 09 2004, pp. 1051–1060.