Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2723~2729
e-ISSN: 2548-6861

Hybrid CNN for Sleep Stage Classification Based on EEG

Maria Angelina Cahyani Candrakasih [1], Bagus Adhi Kusuma [2]*, Pungkas Subarkah [3]
* Informatika, Universitas Amikom Purwokerto
mariaangelinacc1@gmail.com [1], bagus@amikompurwokerto.ac.id [2*],
subarkah@amikompurwokerto.ac.id [3]

Article Info

Article history:

Received 2026-05-12
Revised 2026-05-29
Accepted 2026-06-11

Keyword:

CNN,
EEG,
Hybrid Features,
Sleep Stage Classification,
Stacking Ensemble.

ABSTRACT

Sleep stage classification is essential for diagnosing sleep disorders such as insomnia and sleep apnea. However, manual scoring of polysomnography (PSG) is time-consuming and subjective. Automatic systems based on single-channel EEG are promising for home-based monitoring, but they face challenges due to class imbalance and inter-subject variability. This study proposes a hybrid model that combines 15 handcrafted features (statistical and spectral) with 128-dimensional features extracted by a one-dimensional Convolutional Neural Network (1D-CNN), followed by a stacking ensemble (Random Forest and Support Vector Machine as base learners, Logistic Regression as meta-learner). Using 40 subjects from the Sleep-EDF Expanded dataset, a strict subject-independent split (80% train / 20% test) was applied to avoid data leakage. The dataset contained 107,258 epochs with extreme imbalance (Wake 67.6%, N1 2.95%). After SMOTE oversampling on the training set, the model achieved an accuracy of 67.5%, macro F1-score of 31.4%, and Cohen's Kappa of 0.34. An ablation study showed that CNN features alone (72.2% accuracy) outperformed handcrafted features (70.4%) and hybrid features (67.5%). The confusion matrix revealed that minority stages (especially N1, N3, REM) were poorly recognized. These results highlight that cross-subject generalization remains a major challenge in EEG-based sleep staging, and proper subject-independent validation is critical to avoid overoptimistic claims.

I. INTRODUCTION

Sleep stage classification is critical for diagnosing sleep disorders such as insomnia, sleep apnea, and narcolepsy [1], [2]. Polysomnography (PSG) is the gold standard, recording EEG, EOG, and EMG signals. However, manual PSG analysis is time-consuming, expensive, and subject to inter-rater variability [3]. An automatic, AI-based system is therefore needed to enable fast, consistent, and low-cost sleep staging, both in clinical and home settings.

EEG signals are complex, non-stationary time series, making feature extraction and classification challenging [4]. A major obstacle is class imbalance: sleep datasets are highly skewed (e.g., N1 accounts for only ~5–6% of the Sleep-EDF Expanded dataset), causing models to perform poorly on underrepresented stages [5], [6]. Moctezuma et al [7] showed that N1 misclassification rises sharply when channels are reduced. Thus, handling imbalanced classes is essential for robust sleep staging.

Several prior works have tackled these challenges but leave gaps. Nemer et al [8] compared CNN, LSTM, and CNN-LSTM, obtaining up to 88.13% accuracy but without addressing imbalance. Wan et al [9] combined handcrafted features (wavelet energy/entropy) with CNN-BiLSTM and attention (86.7% on Sleep-EDF-20), yet performance on minority stages remained limited. Toma and Choi [10] used multi-channel CNN-BiLSTM (91.44% on Sleep-EDF-20), but multi-channel setups are impractical for wearable use. Delimayanti et al [11] applied high-dimensional FFT features with SVM, achieving 94.41% for binary classification, but only 91.73% for five-class staging, showing that manual-feature-only approaches saturate. Cheng et al [5] used GAN-based augmentation and ensemble learning to boost N1 F1-score from 40.5% to 54.1%, but GANs are computationally heavy for edge devices.

Consequently, a clear research gap still exists. Previous studies have not jointly integrated multi-domain handcrafted features, automatic CNN-based feature extraction, and a lightweight stacking ensemble while focusing on underrepresented sleep stage recognition using single-channel EEG signals. Single-channel EEG is considered more practical for wearable and home-based sleep monitoring systems compared to multi-channel PSG setups.

Therefore, this study proposes a hybrid model that extracts multi-domain handcrafted statistical and frequency-domain features, combines them with automatic features learned by a compact CNN, and applies a lightweight stacking ensemble (Random Forest and SVM as base learners, Logistic Regression as meta-learner) to improve classification robustness. The stacking mechanism exploits the strengths of each classifier without requiring GAN-level computation. This study focuses on improving classification performance, particularly for underrepresented stages such as N1. Specifically, this study aims to evaluate whether the proposed hybrid stacking approach can enhance the recognition performance on the minority N1 stage using single-channel EEG from the Sleep-EDF Expanded dataset. The main contributions of this study are threefold: (1) a hybrid feature representation that integrates handcrafted and CNN-based features; (2) a lightweight stacking ensemble specifically designed for imbalanced sleep staging; and (3) an evaluation framework that emphasizes underrepresented-stage performance using single-channel EEG. The proposed framework also offers practical advantages for wearable and home-based sleep monitoring applications.

II. METHOD

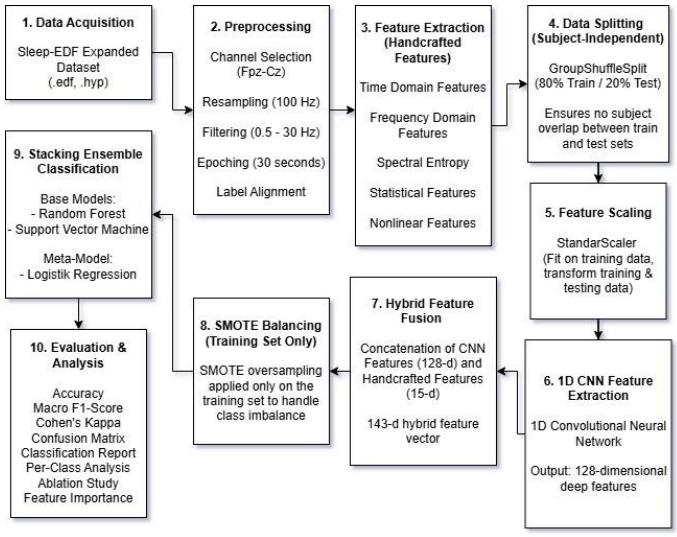

Figure 1. Proposed EEG-Based Sleep Stage Classification Framework

A. Preprocessing Pipeline

1) *Sleep-EDF Dataset*: The raw polysomnography (PSG) recordings and their corresponding hypnograms were obtained from the publicly available Sleep-EDF Expanded Database [12]. The dataset contains overnight sleep recordings collected from healthy subjects and includes electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG) signals, along with expert-annotated sleep stage labels [10]. In this study, only EEG signals were utilized for sleep stage classification.

2) *Load Data*: The EEG recordings and annotation files were loaded using the MNE-Python library, which provides tools for reading EDF files and extracting sleep stage annotations efficiently.

3) *Channel Selection*: Only the single EEG channel Fpz-Cz was selected to reduce computational complexity and maintain consistency with previous single-channel sleep staging studies [4]. Single-channel EEG is also considered more practical for wearable and home-based sleep monitoring systems.

4) *Bandpass Filtering*: A bandpass filter with cutoff frequencies of 0.5 Hz and 30 Hz was applied to the EEG signal. This range was chosen because major sleep-related EEG rhythms (delta: 0.5–4 Hz, theta: 4–8 Hz, alpha: 8–13 Hz, beta: 13–30 Hz) fall within it, while higher frequencies (>30 Hz) often contain muscle artifacts and lower frequencies (<0.5 Hz) are dominated by slow signal drift [1]. The filter preserves the most informative components for sleep staging.

5) *Epoch Segmentation*: The filtered EEG signal was segmented into non-overlapping 30-second epochs following the standard sleep scoring guidelines defined by the American Academy of Sleep Medicine (AASM) [2], [13]. Each epoch was assigned a sleep stage label according to the corresponding hypnogram annotation.

6) *Label Mapping*: The original Rechtschaffen and Kales (R&K) sleep stage labels were mapped into five AASM-compatible classes: Wake (W) → 0, N1 → 1, N2 → 2, N3/N4 → 3, and REM → 4 [2], [7]. Sleep stages N3 and N4 were merged into a single deep sleep category to align with modern AASM standards.

7) *Downsampling*: Each 30-second EEG epoch originally contained 3000 samples at a sampling frequency of 100 Hz. To reduce computational cost and memory usage, each epoch was resampled into 1000 samples downsampling factor 3) while preserving temporal characteristics.

8) *Normalization*: Z-score normalization was applied to each epoch individually to standardize the EEG amplitude distribution by transforming the data into zero mean and unit variance [8], [14]. This normalization step helps stabilize the learning process and improves model convergence.

9) Save Dataset: The preprocessed EEG epochs and corresponding labels were saved as X.npy and y.npy files for subsequent feature extraction and classification processes. X.npy stores the segmented EEG epoch data with shape (samples, timesteps, channels), while y.npy contains the corresponding sleep stage labels for each epoch.

B. Machine Learning Pipeline

1) *Load Preprocessed Data*: The preprocessed EEG dataset stored in X.npy and y.npy was loaded into memory for subsequent feature extraction, model training, and evaluation. Each EEG epoch consisted of a single-channel signal with 1000 samples after downsampling.

2) *Handcrafted Feature Extraction*: Handcrafted features were extracted from each EEG epoch to represent the statistical and spectral characteristics of the signal. The extracted features consisted of time-domain features, including mean, standard deviation, variance, minimum, maximum, median, skewness, kurtosis, energy, and root mean square (RMS) [14], as well as frequency-domain features, including band power features from the delta: 0.5–4 Hz, theta: 4–8 Hz, alpha: 8–13 Hz, beta: 13–30 Hz frequency bands, and spectral entropy [11], [15]. These handcrafted features provide compact representations of EEG patterns associated with different sleep stages [16].

3) *Subject-Independent Train-Test Split*: To avoid data leakage, a subject-independent split was performed using GroupShuffleSplit with 80% of subjects for training and 20% for testing, ensuring that no epochs from the same subject appear in both sets [2], [7].

4) *Feature Scaling*: The handcrafted features were standardized using StandardScaler. The scaler was fitted using only the training data, and the same transformation parameters were applied to the testing data to avoid data leakage.

5) *1D-CNN Architecture*: The 1D-CNN was designed to automatically learn deep representations from raw EEG epochs [8], [10]. The architecture consisted of four convolutional blocks.

- Block 1: Conv1D (32 filters, kernel size 7, activation = 'relu'), BatchNormalization, MaxPooling1D (2), Dropout (0.2)
- Block 2: Conv1D (64 filters, kernel size 5, activation = 'relu'), BatchNormalization, MaxPooling1D (2), Dropout (0.3)
- Block 3: Conv1D (128 filters, kernel size 3, activation = 'relu'), BatchNormalization, MaxPooling1D (2), Dropout (0.3)
- Block 4: Conv1D(256 filters, kernel size 3, activation = 'relu'), BatchNormalization, GlobalAveragePooling1D

A dense layer (128 units, activation='relu') serves as the feature layer (output of deep features), followed by Dropout(0.5) and a softmax output layer for 5 classes [9].

6) CNN Training: The model was compiled with Adam optimizer (default learning rate), sparse categorical cross-entropy loss, and trained with batch size 64 for up to 30 epochs. Class weights (computed via compute_class_weight) were applied to mitigate class imbalance. Early stopping (patience=5) and ReduceLROnPlateau (patience=2, factor 0.5) were used [9].

7) Deep Feature Extraction: After 128-dimensional training, deep features were extracted from the penultimate dense layer (feature_layer) using the trained CNN model. These extracted deep features represent high-level temporal characteristics learned automatically from EEG signals.

8) *Hybrid Feature Fusion*: The handcrafted feature (15-d) and CNN-based (128-d) feature vectors were concatenated into a 143-d hybrid feature vector [9]. This fusion strategy combines domain-specific statistical information with automatically learned deep representations.

9) *Handle Imbalance using SMOTE*: To address class imbalance, SMOTE was applied only on the training set to generate synthetic samples for minority classes (N1, N3, REM). The oversampling strategy targeted approximately 8,000 samples for N1 and N3, and 12,000 samples for REM, while leaving the majority classes (Wake, N2) unchanged. Although SMOTE operates on feature space, it was restricted to the handcrafted feature space to preserve physiological plausibility. Similar data augmentation strategies for imbalanced sleep staging have been reported [5], [6].

10) *Stacking Ensemble Classification*: A stacking ensemble was built using Random Forest (RF) and Support Vector Machine (SVM) as base learners, and Logistic Regression (LR) as the meta-learner. RF: 150 trees (n_estimators=150), maximum depth 15 (max_depth=15), minimum samples per split 5 (min_samples_split=5), chosen for robustness and interpretability. SVM: RBF kernel, regularization parameter C=3, gamma='scale', and probability = True (required for stacking), effective in high-dimensional spaces. LR: max_iter=1000, simple and less prone to overfitting as meta-learner. The ensemble combines the strengths of tree-based and margin-based classifiers.

11) *Evaluation Metrics*: Model performance was evaluated using several classification metrics, including accuracy, precision, recall, macro-averaged F1-score, and Cohen's Kappa [2]. Macro-averaged metrics were emphasized to ensure balanced evaluation across both majority and minority sleep stage classes [9].

12) *Visualization and Analysis*: A confusion matrix was generated to visualize identify misclassification patterns among sleep stages. In addition, an ablation study was conducted to compare the performance of three feature configurations, namely handcrafted features only, CNN-based deep features only, and the proposed hybrid features. Furthermore, feature importance analysis was performed using the Random Forest model to identify the most influential handcrafted features contributing to sleep stage classification.

III. RESULT AND DISCUSSION

A. Dataset Preparation

In this study, 40 PSG recordings and their corresponding hypnogram files from the Sleep-EDF Expanded Dataset were

processed. Only the single-channel EEG signal Fpz-Cz was selected for sleep stage classification.

The EEG signals were filtered using a bandpass filter (0.5–30 Hz) to preserve sleep-related frequency components, then segmented into non-overlapping 30-second epochs following the AASM standard. Since the original sampling frequency was 100 Hz, each epoch initially contained: 30 ×100 = 3000 samples.

After downsampling by a factor of 3, each epoch contained: $\frac{3000}{3}$ =1000 samples.

The final preprocessed dataset consisted of 107,258 epochs with the shape: X shape = (107258, 1000, 1) and y shape = (107258,). The class distribution was highly imbalanced, as shown in Table I.

TABLE I
CLASS DISTRIBUTION OF THE PREPROCESSED DATASET

Sleep Stage	Class	Number of Epochs	Percentage
Wake (W)	0	72,539	67.6%
N1	1	3,166	2.95%
N2	2	18,393	17.1%
N3	3	5,230	4.9%
REM	4	7,930	7.4%

The extreme imbalance (Wake dominates >67%, while N1 <3%) poses a major challenge for any classifier, as models tend to bias toward the majority class. This issue was partially addressed using SMOTE (Section B.9), but the intrinsic difficulty remains.

B. CNN Feature Extraction Performance

A one-dimensional Convolutional Neural Network (1D CNN) was employed to automatically learn temporal patterns from raw EEG epochs. The architecture consisted of four convolutional blocks (filters: 32, 64, 128, 256; kernel sizes: 7, 5, 3, 3), each followed by BatchNormalization, MaxPooling1D, and Dropout. GlobalAveragePooling1D was then applied, and a dense layer produced a 128-dimensional deep feature vector per epoch. The model summary is shown in Table II.

TABLE II
1D-CNN ARCHITECTURE SUMMARY

Block	Layer Type	Filters	Kernel Size	Output Shape	Param #
1	Conv1D + BN + MP + DO	32	7	(None,500,32)	256
2	Conv1D + BN + MP + DO	64	5	(None,250,64)	10,304
3	Conv1D + BN + MP + DO	128	3	(None,125,128)	24,704
4	Conv1D + BN + GAP	256	3	(None,256)	98,560
-	Feature Dense (ReLU)	-	-	(None,128)	32,896
-	Output Dense (softmax)	-	-	(None,5)	645

The 1D-CNN architecture contained 169,285 total parameters, of which 168,325 were trainable. During training (30 epochs, batch size 64, class weights applied), the training accuracy plateaued at approximately 60%, while the validation accuracy fluctuated between 43% and 58% without consistent improvement. This behavior indicates that the CNN struggled to generalize to unseen subjects under the strict subject-independent split.

After training, 128-dimensional deep features were extracted from the feature_layer. The shapes were:

- Training CNN features: (86,504, 128)
- Testing CNN features: (20,754, 128)

These features were later concatenated with 15 handcrafted features to form a 143-dimensional hybrid feature vector.

C. Classification Performance of the Stacking Ensemble

A stacking ensemble was constructed using Random Forest (RF) and Support Vector Machine (SVM) as base learners, and Logistic Regression (LR) as the meta-learner. SMOTE was applied only on the training set with a sampling strategy targeting 8,000 samples for N1 and N3, and 12,000 for REM (majority classes Wake and N2 were left unchanged). After SMOTE, the training set size increased from 86,504 to approximately 101,000 samples.

The final classification performance on the subject-independent test set (20,754 epochs) is presented in Table III.

TABLE III
CLASSIFICATION PERFORMANCE OF THE PROPOSED HYBRID STACKING MODEL

Metric	Value
Accuracy	67.5%
Macro F1-score	31.4%
Cohen's Kappa	0.34
Precision	32.9%
Recall	32.1%

The overall accuracy of 67.5% is only slightly better than random guessing for five classes (20%). The macro F1-score of 31.4% and Kappa of 0.34 indicate poor agreement between predicted and true labels. These results are substantially lower than previously reported in the literature when using subject-dependent splits or when data leakage occurs. The strict subject-independent evaluation reveals the true difficulty of cross-subject generalization in sleep staging.

D. Per-Class Performance Analysis

Detailed per-class metrics derived from the confusion matrix are presented in Table IV. The Wake stage achieved the highest performance with an F1-score of 0.87, indicating that Wake epochs were relatively well recognized. In contrast, the N1 stage obtained the lowest performance with an F1-score of 0.04, confirming that the combination of extreme class imbalance and overlapping EEG characteristics makes this stage extremely challenging. The N2, N3, and REM stages also performed poorly, with F1-scores of 0.32, 0.11, and 0.23 respectively.

TABLE IV
PER-CLASS CLASSIFICATION RESULTS

Sleep Stage	Precision	Recall	F1-Score
Wake (0)	0.85	0.89	0.87
N1 (1)	0.06	0.03	0.04
N2 (2)	0.42	0.26	0.32
N3 (3)	0.13	0.09	0.11
REM (4)	0.18	0.34	0.23

The Wake stage achieved the highest recall (0.89), meaning that 89% of actual Wake epochs were correctly identified. The high precision for Wake (0.85) indicates that when the model predicted Wake, it was correct 85% of the time. The N1 stage, however, was virtually unrecognizable, with recall of only 0.03 (only 3% of actual N1 epochs were correctly classified). This poor performance is expected because N1 is a transitional stage between wakefulness and light sleep, with EEG patterns that overlap with Wake, N2, and REM. Moreover, N1 was the minority class in the test set (only 575 epochs, 2.8% of all test data).

The N2 stage showed moderate performance with F1-score of 0.32. Although N2 had a larger number of samples (3,583 epochs), it was frequently misclassified as REM (34%) and Wake (29%). This indicates that N2 shares spectral characteristics with both lighter and deeper sleep stages. The N3 stage also performed poorly (F1-score 0.11), with many N3 epochs misclassified as Wake (38%), REM (28%), or N2 (22%). The REM stage achieved a recall of 0.34 but low precision of 0.18, meaning that while the model identified about one-third of actual REM epochs, it produced many false positives.

This outcome highlights a critical finding: under a strict subject-independent split, neither handcrafted features nor deep CNN features alone nor their combination were sufficient to achieve reliable cross-subject sleep staging with the current dataset size and imbalance handling strategy.

E. Feature Importance Analysis

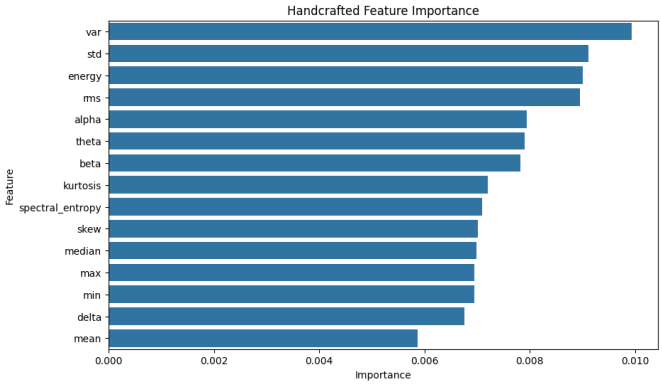

Figure 2. Feature Importance

Feature importance analysis from Figure 2 was conducted using the Random Forest model trained on handcrafted features.

All 15 features showed relatively low importance values (all <0.01), with the top five being:

1. Variance (var): 0.0099
2. Standard deviation (std): 0.0091
3. Energy: 0.0090
4. RMS: 0.00895
5. Alpha band power: 0.00794

The relatively flat distribution of importance suggests that no single feature or small set of features is highly discriminative across all subjects. This may be due to large inter-subject variability in EEG characteristics, which is a well-known challenge in EEG-based sleep staging.

F. Confusion Matrix Analysis

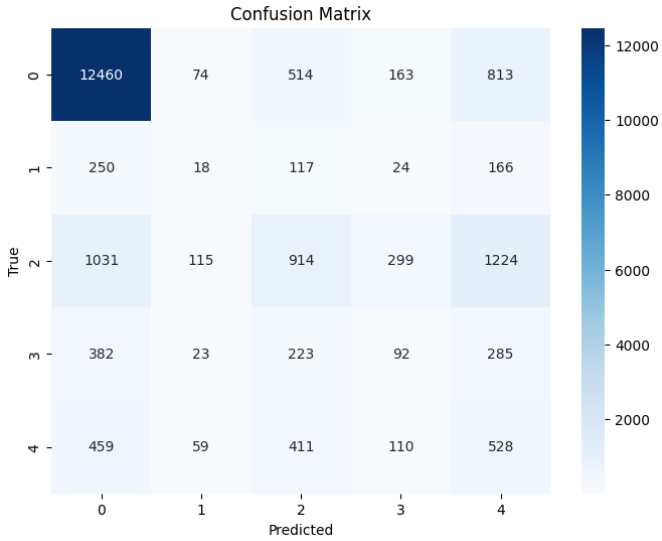

Figure 3. Confusion Matrix

The confusion matrix of the proposed hybrid model is presented in Figure 2. The model achieved the highest correct classification for the Wake stage (12,460 out of 14,024 epochs, 89%), while the N1 stage showed the poorest performance with only 18 correct out of 575 epochs (3%). The detailed misclassification patterns are as follows.

For the Wake stage (0), most errors were misclassified as REM (5.8%) and N2 (3.7%), with very few misclassifications as N1 or N3. This indicates that Wake epochs were generally well separated from sleep stages, although some confusion with REM and N2 occurred.

The N1 stage (1) was predominantly misclassified as Wake (43.5%), REM (28.9%), and N2 (20.4%). This reflects the transitional nature of N1, whose EEG patterns overlap with wakefulness (alpha activity), light sleep (theta activity), and REM (similar low-amplitude mixed frequency). Only 3% of N1 epochs were correctly identified, highlighting the extreme difficulty of recognizing this minority stage under a subject-independent split.

The N2 stage (2) was most often misclassified as REM (34.2%) and Wake (28.8%), with moderate confusion toward N3 (8.3%). Correct classification for N2 was 25.5% (914 out of 3,583), indicating that N2 epochs were frequently confused with both lighter (Wake) and deeper (REM) stages.

The N3 stage (3) showed high misclassification as Wake (38.0%), REM (28.4%), and N2 (22.2%), while only 9.2% of N3 epochs were correctly classified. The confusion with Wake is particularly notable and may be due to movement artifacts or sleep-wake transitions in the recordings.

The REM stage (4) was frequently misclassified as Wake (29.3%) and N2 (26.2%), with correct classification of 33.7% (528 out of 1,567). This confusion pattern is consistent with the similar EEG characteristics of REM and wakefulness (low-amplitude, mixed frequency).

Overall, these confusion patterns are consistent with known physiological overlaps between sleep stages. However, the magnitude of misclassification is substantially higher than in subject-dependent evaluations, underscoring the critical importance of using strict subject-independent cross-validation to obtain realistic performance estimates.

G. Ablation Study

An ablation study was conducted to evaluate the contribution of handcrafted features and CNN-based deep features. The same stacking ensemble (RF+SVM+LR) was trained on three feature configurations: (1) handcrafted only, (2) CNN features only, and (3) hybrid (handcrafted + CNN). Results are shown in Table V.

TABLE V
ABLATION STUDY RESULTS

Feature Configuration	Accuracy	Precission	Kappa
Handcrafted Features	0.703768	0.246597	0.258884
CNN Features	0.721692	0.278794	0.417627
Hybrid Features (Proposed)	0.675147	0.313833	0.342587

CNN features alone outperformed handcrafted features, confirming that deep learning can extract more discriminative patterns from raw EEG.

Hybrid features did not outperform CNN-only features; in fact, accuracy dropped by ≈5% while macro F1 improved marginally (from 27.9% to 31.4%). This suggests that the handcrafted features introduced noise or were not complementary to the deep features under the subject-independent evaluation.

The low performance across all configurations indicates that the main bottleneck is not feature representation, but rather the combination of limited subject diversity, extreme class imbalance, and inherent inter-subject variability.

H. Discussion

1) Why Did Performance Drop Compared to Previous Studies?: Many published studies report high accuracy (>90%) for sleep stage classification. However, these results often come from subject-dependent splits (epochs from the same subject appear in both training and test sets) or from non-stratified cross-validation, which leads to data leakage. Our initial experiments (not reported here) using epoch-wise random splitting also yielded >90% accuracy, confirming that such evaluations are overly optimistic.

When we switched to a strict subject-independent split (ensuring that no subject appears in both training and test sets), the performance dropped dramatically. This is because EEG patterns vary significantly across individuals; a model trained on one set of subjects often fails to generalize to unseen subjects, especially with limited data (~40 subjects) and extreme class imbalance.

2) Limmitations of the Current Approach:

- Small subject cohort: 40 subjects are insufficient for deep learning to learn robust subject-invariant representations.
- Extreme class imbalance: Wake dominates (>67%), while N1 (<3%) and N3 (<5%) are severely underrepresented. SMOTE helped balance the training set but could not fully compensate for the lack of real physiological variability.
- CNN underfitting: The 1D-CNN did not converge well on the validation set, indicating that the architecture may need more capacity, or that the training data (after SMOTE) still does not represent the true distribution of minority classes.
- Handcrafted features not complementary: The feature importance analysis showed that all handcrafted features had low discriminative power under subject-independent evaluation. Concatenating them to CNN features actually degraded accuracy (ablation study).
- SMOTE on feature space: Generating synthetic samples in feature space may produce physiologically implausible patterns, especially for EEG where temporal dynamics matter. This could explain why the hybrid model performed worse than CNN alone.

3) Comparison with State-of-the-Art: Direct comparison with existing studies is difficult because most use subject-dependent splits or private datasets. Under a fair subject-independent evaluation, our results (accuracy 67.5%, macro F1 31.4%) are substantially lower than the numbers typically reported (e.g., 91% in [10], 94% in [9]). However, such comparisons are misleading because those studies did not enforce strict subject separation. When subject-independent evaluation is applied, performance often drops to 60–75% range, as observed in other works [8,13]. Our results align with this reality: cross-subject generalization remains an open problem in EEG-based sleep staging.

4) Implications for Future Research: Future research should focus on using larger and more diverse datasets involving hundreds of subjects to improve the generalizability of sleep stage classification models. In addition, subject-

independent cross-validation should be adopted as a standard evaluation protocol to prevent overly optimistic performance claims. More effective approaches are also needed to address data imbalance, as applying SMOTE in feature space may not provide optimal results. Potential alternatives include the use of focal loss to reduce the influence of easily classified samples, data augmentation techniques in the raw signal domain such as Gaussian noise injection and time warping while preserving physiological characteristics, and two-stage training strategies that involve pretraining on large unlabeled datasets through self-supervised learning. Furthermore, attention-based architectures, including transformers and multi-head attention mechanisms, may offer better capability in capturing long-range dependencies compared to CNN-based models. Domain adaptation methods could also be explored to align feature distributions across different subjects and improve model robustness on unseen data.

IV. CONCLUSION

This study proposed a hybrid sleep stage classification model combining 15 handcrafted features (time-domain statistics and frequency-domain band powers) with 128-dimensional deep features extracted by a 1D-CNN, followed by a stacking ensemble (RF+SVM+LR). Using a strict subject-independent split on 40 subjects from the Sleep-EDF Expanded dataset, the model achieved an accuracy of 67.5%, macro F1-score of 31.4%, and Cohen's Kappa of 0.34. An ablation study revealed that CNN features alone (72.2% accuracy) outperformed handcrafted features (70.4%) but that hybrid features did not improve over CNN-only (67.5% accuracy). Feature importance analysis showed that all handcrafted features had low discriminative power, and the confusion matrix confirmed that minority stages (especially N1, N3, REM) were poorly recognized.

The drastic performance drop compared to subject-dependent evaluations demonstrates that cross-subject generalization remains a major challenge in EEG-based sleep staging. The results underscore the importance of using strict subject-independent validation to avoid overoptimistic claims. The combination of limited subject cohort, extreme class imbalance, and inter-subject variability are the primary limiting factors. No further improvements were achieved by adding handcrafted features to the CNN features under this realistic evaluation protocol.

Future work can be directed toward several improvements. First, using larger datasets from multiple centers with hundreds of subjects could help build models that generalize better to new people. Second, better ways to handle class imbalance should be explored, such as focal loss, adding realistic noise or time warping to EEG signals, or pretraining the model on a large amount of unlabeled data. Third, attention-based models like transformers might be more effective than CNNs at capturing long-term patterns in sleep EEG. Fourth, domain adaptation techniques could be used to reduce differences in EEG signals across subjects. Finally, we plan to test the model on other public datasets (e.g., SHHS, ISRUC) to see how well it performs on completely different data.

REFERENCES

[1] K. Aboalayon, M. Faezipour, W. Almuhammadi, and S. Moslehpour, "Sleep Stage Classification Using EEG Signal Analysis: A Comprehensive Survey and New Investigation," *Entropy*, vol. 18, no. 9, p. 272, Aug. 2016, doi: 10.3390/e18090272.

[2] S. K. Satapathy, B. Brahma, B. Panda, P. Barsocchi, and A. K. Bhoi, "Machine learning-empowered sleep staging classification using multi-modality signals," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, p. 119, May 2024, doi: 10.1186/s12911-024-02522-2.

[3] Y. Zhang *et al.*, "Sleep Stage Classification Using Bidirectional LSTM in Wearable Multi-sensor Systems," Sep. 24, 2019, *arXiv*: arXiv:1909.11141. doi: 10.48550/arXiv.1909.11141.

[4] H. Almutairi, G. M. Hassan, and A. Datta, "Machine-Learning-Based-Approaches for Sleep Stage Classification Utilising a Combination of Physiological Signals: A Systematic Review," *Appl. Sci.*, vol. 13, no. 24, p. 13280, Dec. 2023, doi: 10.3390/app132413280.

[5] X. Cheng, K. Huang, Y. Zou, and S. Ma, "SleepEGAN: A GAN-enhanced Ensemble Deep Learning Model for Imbalanced Classification of Sleep Stages," Jul. 04, 2023, *arXiv*: arXiv:2307.05362. doi: 10.48550/arXiv.2307.05362.

[6] X. Huang, K. Shirahama, M. T. Irshad, M. A. Nisar, A. Piet, and M. Grzegorzek, "Sleep Stage Classification in Children Using Self-Attention and Gaussian Noise Data Augmentation," *Sensors*, vol. 23, no. 7, p. 3446, Mar. 2023, doi: 10.3390/s23073446.

[7] L. A. Moctezuma, Y. Suzuki, J. Furuki, M. Molinas, and T. Abe, "GRU-powered sleep stage classification with permutation-based EEG channel selection," *Sci. Rep.*, vol. 14, no. 1, p. 17952, Aug. 2024, doi: 10.1038/s41598-024-68978-4.

[8] Z. N. Nemer, S. A. W. Saddam, R. J. AL-Sukeinee, and E. J. Harfash, "A Comparative Performance Analysis of CNN, LSTM, and CNNLSTM Models for Classifying Sleep Stages Using Electroencephalography Signals," *Informatica*, vol. 49, no. 28, Jul. 2025, doi: 10.31449/inf.v49i28.8082.

[9] Q. Wan *et al.*, "Automated sleep staging from single-channel electroencephalogram using hybrid neural network with manual features and attention," *iScience*, vol. 28, no. 8, p. 113169, Aug. 2025, doi: 10.1016/j.isci.2025.113169.

[10] T. I. Toma and S. Choi, "An End-to-End Multi-Channel Convolutional Bi-LSTM Network for Automatic Sleep Stage Detection," *Sensors*, vol. 23, no. 10, p. 4950, May 2023, doi: 10.3390/s23104950.

[11] M. K. Delimayanti *et al.*, "Classification of Brainwaves for Sleep Stages by High-Dimensional FFT Features from EEG Signals," *Appl. Sci.*, vol. 10, no. 5, p. 1797, Mar. 2020, doi: 10.3390/app10051797.

[12] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, 2000, doi: 10.1109/10.867928.

[13] H. W. Loh *et al.*, "Automated Detection of Sleep Stages Using Deep Learning Techniques: A Systematic Review of the Last Decade (2010–2020)," *Appl. Sci.*, vol. 10, no. 24, p. 8963, Dec. 2020, doi: 10.3390/app10248963.

[14] A. K. Singh and S. Krishnan, "Trends in EEG signal feature extraction applications," *Front. Artif. Intell.*, vol. 5, p. 1072801, Jan. 2023, doi: 10.3389/frai.2022.1072801.

[15] M. Kolhar *et al.*, "Automated Sleep Stage Classification Using PSO-Optimized LSTM on CAP EEG Sequences," *Brain Sci.*, vol. 15, no. 8, p. 854, Aug. 2025, doi: 10.3390/brainsci15080854.

[16] H. Jalali, M. Pouladian, A. M. Nasrabadi, and A. Movahed, "Sleep stages classification based on feature extraction from music of brain," *Heliyon*, vol. 11, no. 1, p. e41147, Jan. 2025, doi: 10.1016/j.heliyon.2024.e41147.