Classifying Sleep Stages Using Machine Learning and Deep Learning:
A Comparative Analysis

Azim Chyngozhoev 20212003

Department of Biomedical Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea

Sleep stage classification is crucial to understand sleep patterns and it aids in predicting sleep disorders early on, which is one of the main problems: people are too late to diagnose the disorders. In this study, Using the **PhysioNet Sleep-EDF Expanded Dataset**, I evaluate the impact of incorporating temporal context (TC) and Leave-One-Subject-Out (LOSO) cross-validation for generalizability in two main models: **Random Forest (RF)** and **Long Short-Term Memory (LSTM)** networks. The data contains EEG recordings of patients during sleep and PSG recording with corresponding annotations of events. Containing EEG recordings annotated for five sleep stages: **Wake (W), Stage 1, Stage 2, Stage 3/4 (merged), and REM.** There are many analyses already done to classify different sleep stages in EEG, however, certain stages, such as Stage 1 sleep are still very hard to classify due to their transitional nature. RF achieved a mean accuracy of 68% under LOSO, while LSTM with temporal context attained 66%. I discuss the strengths, weaknesses, and potential future directions of these models in sleep stage classification.

1. Introduction

An electroencephalogram (EEG) is a test that measures electrical activity in the brain. This test also is called an EEG. The test uses small, metal discs called electrodes that attach to the scalp, assigned **channels**. Brain cells communicate via electrical impulses, and this activity shows up as wavy lines on an EEG recording. Brain cells are active all the time, even during sleep.

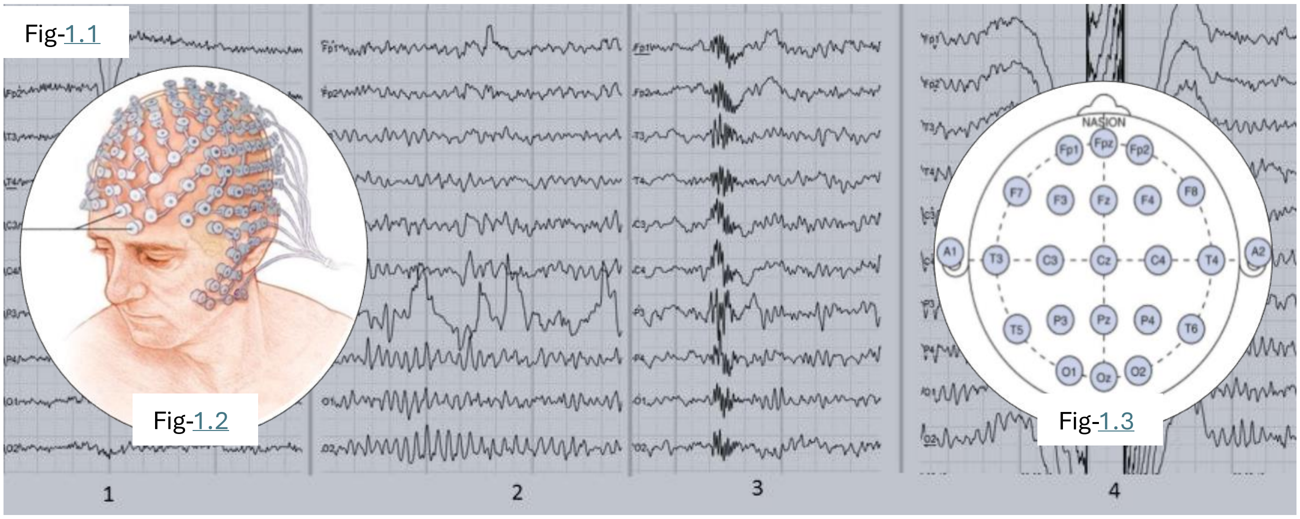


Fig-1.1

Fig-1.2

Fig-1.3

Sleep is a dynamic physiological process consisting of distinct stages: Wake (W), NREM (Stages 1–3/4), and REM. It is important to accurately assess the data and diagnose sleep disorders. However, traditional ways of assessing sleep disorders come with numbers of limitations:

1. It takes time.
   Only certified professionals are able to assess the EEG of the patient and it takes time to record the patients sleep and takes time to assess the data by professionals only.

2. It is subject to professionals abilities.
   While professionals are the best people to address the issues, monitoring and locating early issues with sleep eeg data is difficult and the verdict may vary among professionals.

Thus, implementing AI that learns on many accessible data to accurately monitor brain activies is crucial.

The analysis of **electroencephalogram (EEG)** signals forms the backbone of sleep stage classification. However, there are limitations: classifying **transitional stages (e.g., Stage 1)**, which often overlap with neighboring stages, and ensuring **generalizability** across participants.

This study aims to:

1. Compare the performance of **Random Forest (RF)** and **Long Short-Term Memory (LSTM)** networks for sleep stage classification.
2. Investigate the effect of incorporating **temporal context (TC)** on model accuracy.
3. Evaluate models using **LOSO cross-validation**, ensuring generalization across unseen participants.

---

2. Methodology

2.1 Dataset

I used the dataset from paper by Mourtazaev et al. **Age and gender affect different characteristics of slow waves in the sleep EEG.** The dataset is accessible through Physionet's Sleep-EDF Database Expanded. The dataset seems to be widely recognized in EEG research community thanks to it well annotated and high quality EEG recordings, which is ideal resource for school projects allowing less problems dealing with the data and credibility. It contains recordings of patients Polysomnography (PSG), with events annotations, and EEG recording of brain itself. EEG was analyzed quantitatively during 2 nights in 40 females and 34 males aged between 26 to 101 years old.

The dataset provides two primary EEG channels, **Fpz-Cz** and **Pz-Oz**, along with annotations corresponding to five distinct sleep stages based on the **Rechtschaffen and Kales (R&K)** sleep scoring system: **Wake (W): Awake and alert state with prominent alpha activity.**

- **Stage 1 (N1): Light sleep characterized by low-amplitude theta waves.**
- **Stage 2 (N2): Dominant sleep stage with clear features like sleep spindles (12–16 Hz) and K-complexes.**
- **Stage 3/4 (N3): Deep sleep dominated by high-amplitude delta waves (slow waves).**
- **REM (R): Dream state marked by rapid eye movement and mixed frequency EEG patterns.**

The recordings include **30-second epochs** labeled with their respective sleep stages, this allows for a comprehensive time-series analysis. Each epoch is accompanied by auxiliary signals such as electrooculogram (EOG), chin electromyogram (EMG), and airflow, although only EEG channels were utilized in this study.

Key Properties of the Dataset:

- **Sampling Frequency**: The EEG signals were originally sampled at **100 Hz**. It is sufficient, however, it is still too low. I can't downsample the data which may lead to costly training of the model.

- **Number of Participants**: Although there are 74 participants, this study only aims to train and test the models on 10 participants. Which is enough for Leave-One-Subject out cross validation scheme, helping with generalizability, and cost effective for running both models.
- **Class Distribution**: Usually, EEG sleep datasets are imbalanced. Stage 2 (N2) dominates the recordings, for about 50% of total sleep, while Stage 1 (n1) are underpresented. This is one of the difficulties of the project.

This dataset is highly suitable for investigating the potential of machine learning and deep learning methods in sleep stage classification due to its high-quality annotations, well-documented structure, and relevance to clinical and research applications. The inherent challenges, such as imbalanced class distributions and the transitional nature of certain stages, further highlight the need for robust analytical approaches like temporal modeling and cross-validation.

2.2 Data Preprocessing

Preprocessing is a critical step in ensuring the quality and consistency of EEG data for machine learning and deep learning models. For this study, a combination of normalization, segmentation, and feature extraction techniques was applied to prepare the data for analysis. Below are the detailed steps:

Normalization

To address inter-participant variability in EEG signal amplitude, **z-score normalization** was applied to each channel of the EEG data. This method standardized the data by subtracting the mean and dividing by the standard deviation:

$$Xnorm = \frac{X-\mu}{\sigma}$$

where X is the signal amplitude, μ\muμ is the mean, and σ\sigmaσ is the standard deviation.

- Ensured consistent scaling of the input features, which is critical for convergence in machine learning and deep learning models.
- Reduced bias introduced by varying signal amplitudes across participants.

Epoch Segmentation

The EEG recordings were segmented into **30-second epochs**, as provided by the dataset. Each epoch was assigned a corresponding sleep stage label (Wake, Stage 1, Stage 2, Stage 3/4, or REM) based on manual annotations following the Rechtschaffen and Kales (R&K) scoring guidelines.

- Aligned with standard sleep scoring practices, enabling direct comparison with clinical annotations.
- Served as the primary unit of analysis for both Random Forest and LSTM models.

Band-Pass Filtering

To isolate the frequencies relevant for sleep stage classification, a **band-pass filter** with a range of **0.5–30 Hz** was applied to the EEG signals. This range was chosen because:

- Frequencies below 0.5 Hz represent slow drifts and are often noise artifacts.
- Frequencies above 30 Hz are typically associated with muscle artifacts or environmental noise.

Retained key EEG frequency bands critical for sleep stage classification:

- **Delta (0.5–4 Hz)**: Dominates deep sleep (Stage N3).

- **Theta (4–8 Hz)**: Associated with light sleep (Stages N1 and N2).
- **Alpha (8–12 Hz)**: Prominent during wakefulness.
- **Sigma (12–16 Hz)**: Reflects sleep spindles, critical for Stage N2.
- **Beta (15–30 Hz)**: May occur during REM sleep.

Preprocessing by Dataset Providers

The dataset documentation indicates that slow-wave features were precomputed and used for certain analyses. These included:

Slow-Wave Power (SWP):

- SWP measures the power of **delta waves (0.5–4 Hz)** in the EEG signal, which are prominent during deep sleep (Stages N3/4).
- High SWP is a defining characteristic of slow-wave sleep, making it a crucial feature for distinguishing deep sleep stages.

Slow-Wave Continuity Percentage (SWC%):

- SWC% quantifies the proportion of delta waves that are continuous over a given time window.
- This feature reflects the stability of slow-wave sleep and helps differentiate between deep sleep and other stages.

Purpose of Precomputed Features:

- These measures were likely derived to ensure accurate annotations for slow-wave sleep stages.
- While these features were not explicitly used in our models, they contributed to the robustness of the dataset's annotations, particularly for **Stages N3/4**.

2.3 Models

This study focuses on two distinct modeling approaches: a **Random Forest (RF)** model, a traditional machine learning technique, and a **Long Short-Term Memory (LSTM)** network, a deep learning approach tailored for sequential data. These models were evaluated with and without the inclusion of **temporal context (TC)** to explore their effectiveness in classifying sleep stages.

Random Forest (RF)

An ensemble based machine learning algorithm. It constructs multiple decision trees and combines their outputs, thus making predictions. It is a good approach since it is easily interpretable, and well suited for dataset with features like frequency band power.

Implementation Details:

- **Input Features**:

  The RF model used manually extracted features from EEG signals, such as:

Frequency Band Power: Average power in delta, theta, alpha, beta, and sigma bands.

Statistical Features: Mean, variance, and skewness of the signal amplitudes.

These features were derived from each 30-second epoch of EEG data.

- **Temporal Context**:

  In the **RF-TC variant**, features from neighboring epochs were concatenated to include temporal information about the transitions between sleep stages.

Evaluation:

- RF was trained using **Leave-One-Subject-Out (LOSO)** cross-validation, where the model is trained on data from all participants except one, and tested on the excluded participant.
- RF was then trained using LOSO and Temopral Context

Advantages:

- **Simplicity and Interpretability**: Feature importance rankings provided insights into which aspects of the EEG signals were most informative for each sleep stage.
- **Robustness**: RF handles imbalanced datasets relatively well and is less sensitive to noise in the data compared to neural networks.

Challenges:

- RF struggled to capture complex temporal patterns in sequential data, especially in transitions between stages like Stage 1 and Stage 2.

Long Short-Term Memory (LSTM)

LSTM is a type of recurrent neural network (RNN) designed to model sequential data by maintaining **long-term dependencies**. It is particularly effective for time-series problems like EEG analysis, where temporal patterns are crucial for understanding transitions between sleep stages.

Implementation Details:

- **Input Data**:

  The input to the LSTM model was the raw EEG data (normalized) for each epoch. For the **LSTM-TC** variant, data from neighboring epochs was concatenated to form a larger sequence, enabling the model to learn transitions between sleep stages.

- **Network Architecture**:

  The LSTM network consisted of two layers:

  - **First LSTM Layer**: Captures temporal dependencies within the sequence.
  - **Second LSTM Layer**: Processes long-term relationships across epochs.

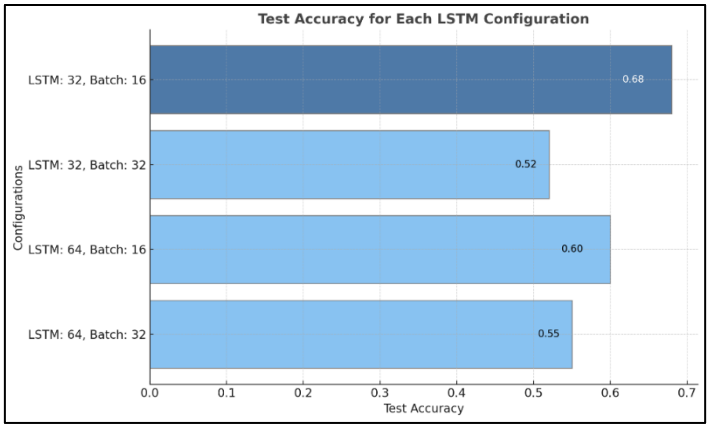
  Dense layers followed the LSTM layers to map the learned representations to the sleep stage classifications.

- **Optimization**:

  In order to not look for best hyperparameters everytime, I implemented Grid hyperparameters that take multiple values for each hyperparameter and run. If the model stop improving, the running stops and switches to next hyperparameter.

  The model was trained using the **Adam optimizer** with categorical cross-entropy as the loss function.

  A batch size of 32 and a learning rate of 0.001 were used for efficient convergence.



- **Training Dynamics**:

Early stopping was implemented to avoid overfitting, based on the validation loss.

Evaluation:

Sine LSTM is a lot more costly than RF, no LOSO was implemented. So there are two variants:

- **Pure LSTM**: Only the current epoch was used as input, limiting the model's ability to handle transitions.
- **LSTM-TC**: Neighboring epochs were included, improving the model's ability to classify transitional stages like Stage 1.

Advantages:

- **Temporal Modeling**: LSTM excels at capturing sequential dependencies, making it better suited for transitional stages than traditional methods like RF.
- **Scalability**: Deep learning models can leverage larger datasets to improve performance further.

Challenges:

- **Data Requirements**: LSTM requires more data to generalize effectively compared to RF, and its performance may be limited on smaller datasets.
- **Complexity**: Training LSTM models is computationally intensive and requires tuning multiple hyperparameters for optimal performance.

---

3. Results

This section presents the results of the Random Forest (RF) and Long Short-Term Memory (LSTM) models applied to the sleep stage classification task. The performance is evaluated using accuracy, confusion matrices, and class-specific metrics such as precision, recall, and F1-score. Both models were analyzed in two configurations: with and without temporal context (TC).

3.1 Random Forest (RF)

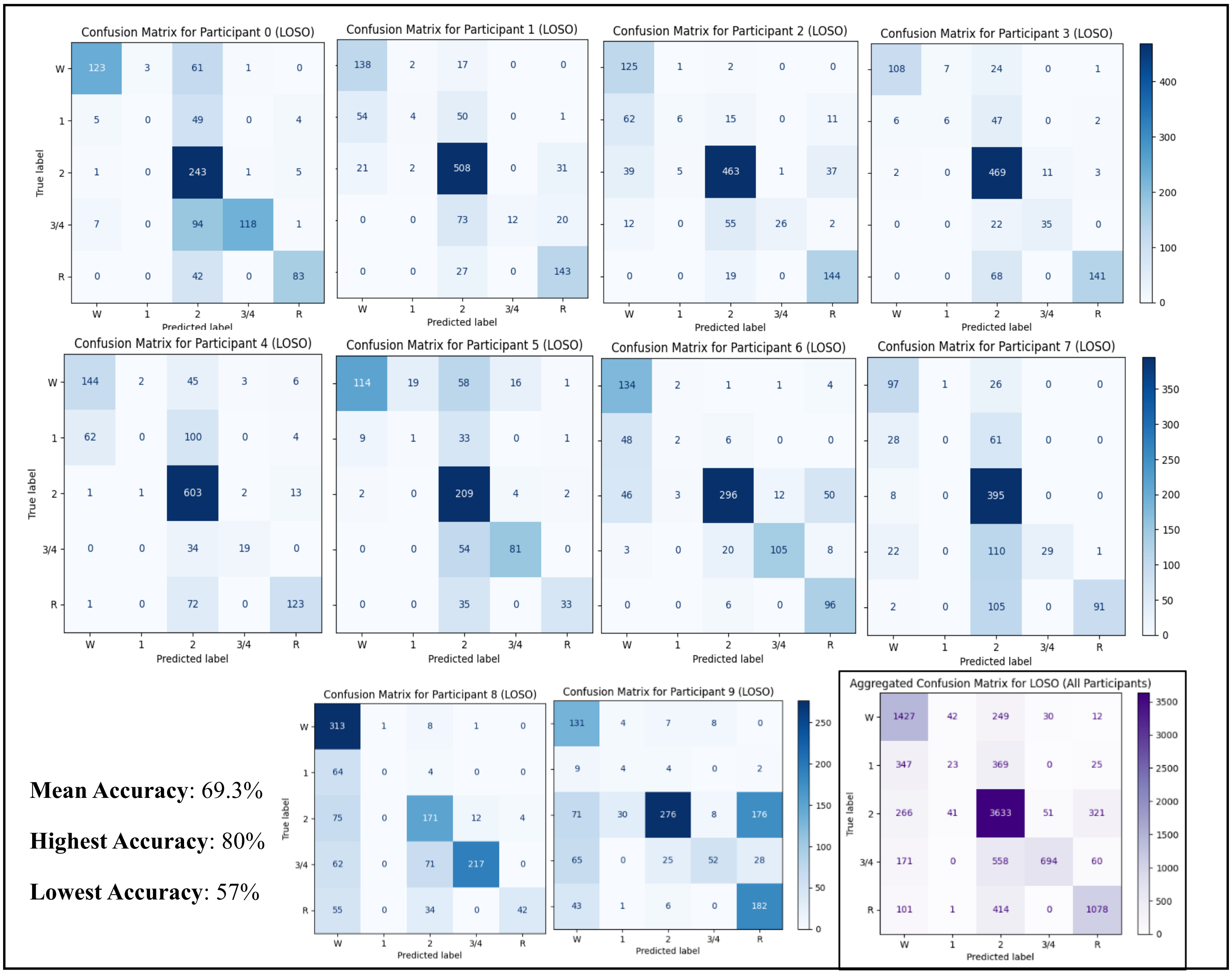
RF without Temporal Context (RF_LOSO)

- **Performance**:

  RF_LOSO achieved a mean accuracy of **68%** across participants when evaluated using Leave-One-Subject-Out (LOSO) cross-validation.

  The model performed best on **Stage 2 (N2)**, the most abundant stage in the dataset, with a high precision of 85%.

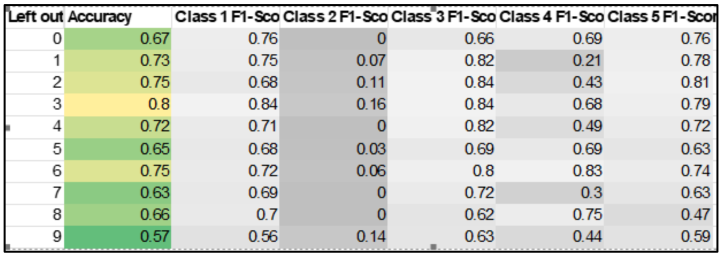
  Performance was weakest for **Stage 1 (N1)**, which was often confused with Wake (W) and Stage 2 (N2), reflecting the transitional nature of N1.



Mean Accuracy: 69.3%

Highest Accuracy: 80%

Lowest Accuracy: 57%

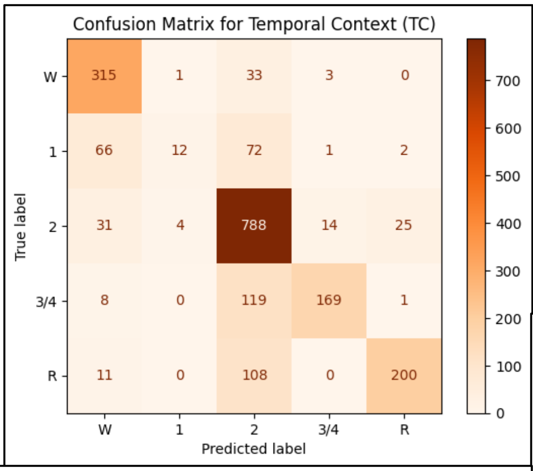


Left out	Accuracy	Class 1 F1-Sco	Class 2 F1-Sco	Class 3 F1-Sco	Class 4 F1-Sco	Class 5 F1-Scor
0	0.67	0.76	0	0.66	0.69	0.76
1	0.73	0.75	0.07	0.82	0.21	0.78
2	0.75	0.68	0.11	0.84	0.43	0.81
3	0.8	0.84	0.16	0.84	0.68	0.79
4	0.72	0.71	0	0.82	0.49	0.72
5	0.65	0.68	0.03	0.69	0.69	0.63
6	0.75	0.72	0.06	0.8	0.83	0.74
7	0.63	0.69	0	0.72	0.3	0.63
8	0.66	0.7	0	0.62	0.75	0.47
9	0.57	0.56	0.14	0.63	0.44	0.59

Confusion Matrix Observations:

- **Stage 2 (N2)** was classified accurately in most cases due to its distinct EEG features, such as spindles and K-complexes.
- **Stage 1 (N1)** suffered from misclassifications due to its low representation in the data and its similarity to adjacent stages.

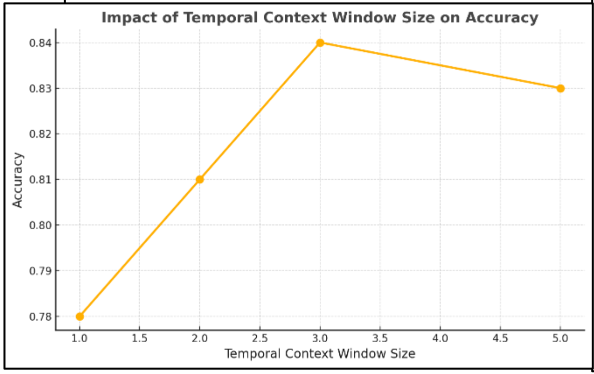
RF with Temporal Context (RF_TC)

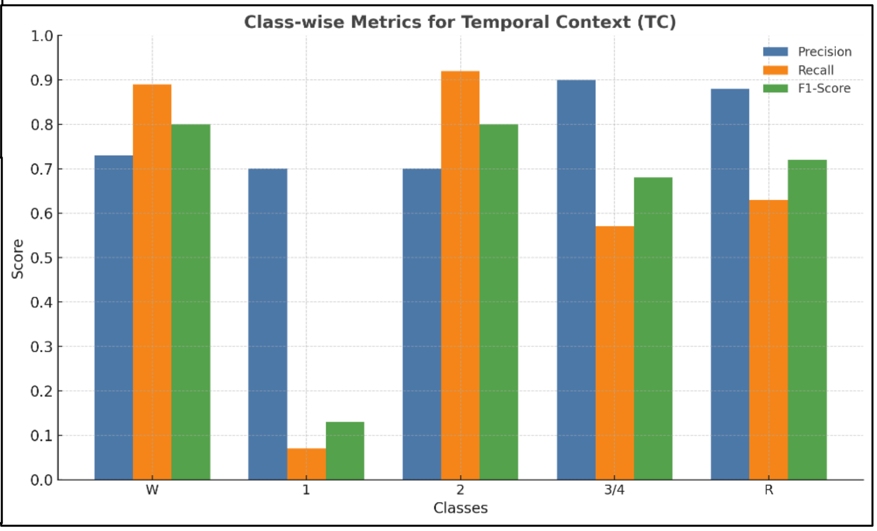


- **Performance**:

Including temporal context improved the accuracy to **76%**, demonstrating the importance of modeling transitions between epochs.

Classification of transitional stages (e.g., N1) improved significantly, with fewer misclassifications into Wake (W).

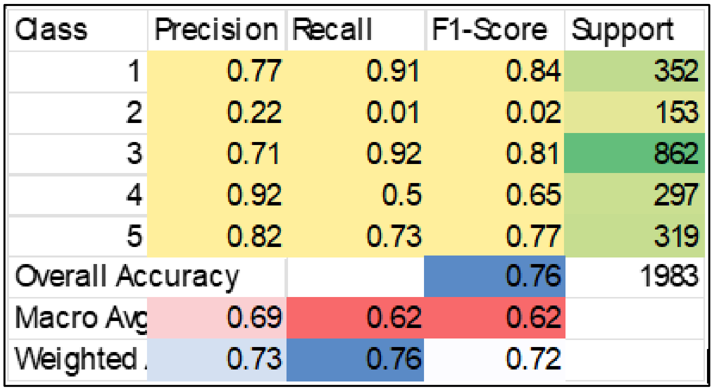




- **Key Observations**:

Temporal context enabled the model to leverage patterns across neighboring epochs, resulting in better identification of sleep stage transitions.

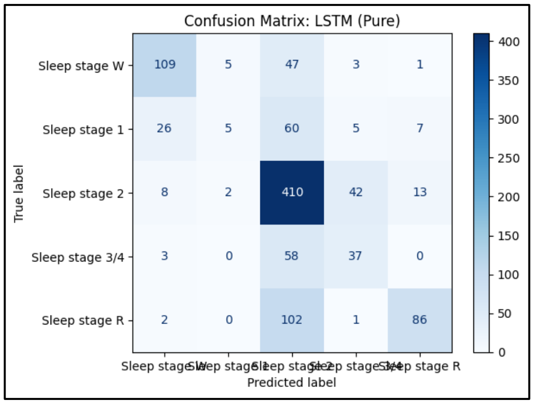
The improvement was most notable for **Stage 1 (N1)** and **REM (R)** stages.



Class	Precision	Recall	F1-Score	Support
1	0.77	0.91	0.84	352
2	0.22	0.01	0.02	153
3	0.71	0.92	0.81	862
4	0.92	0.5	0.65	297
5	0.82	0.73	0.77	319
Overall Accuracy			0.76	1983
Macro Avg	0.69	0.62	0.62	
Weighted .	0.73	0.76	0.72	

3.2 Long Short-Term Memory (LSTM)

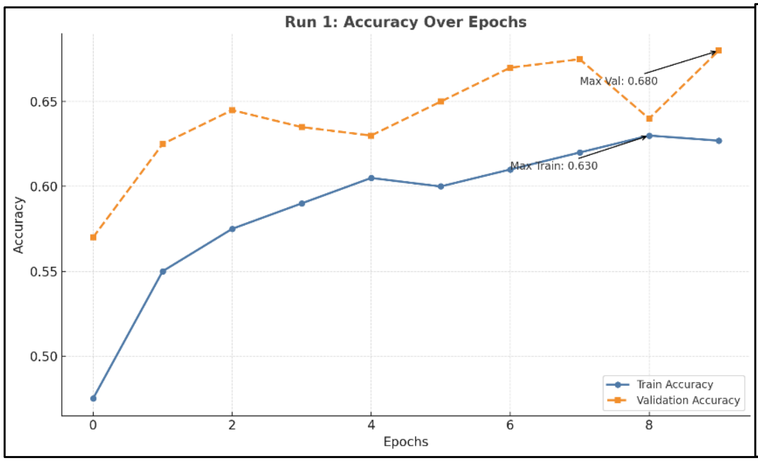
Pure LSTM

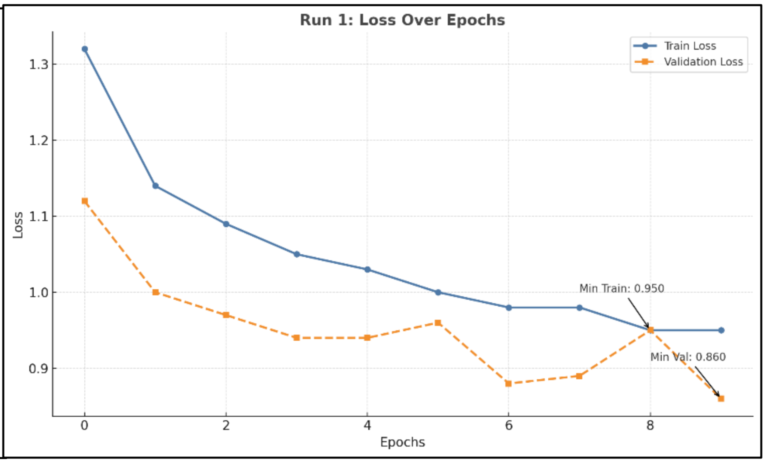


- ***Performance***:

The pure LSTM model achieved an accuracy of **63%**, reflecting its limitations in handling single epochs independently.

Transitional stages like **Stage 1 (N1)** and **REM (R)** were particularly challenging, as these stages require sequential information for accurate classification.



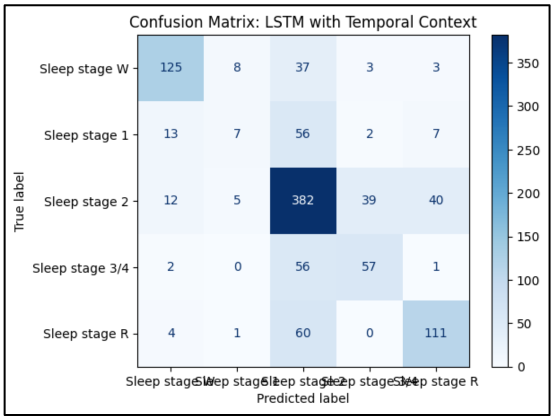


- ***Challenges***:

The model struggled to classify Stage 1 (N1) due to its similarity to adjacent stages.

Unlike RF, LSTM relies heavily on capturing sequential dependencies, which was limited without temporal context.

LSTM with Temporal Context (LSTM_TC)



- ***Performance***:

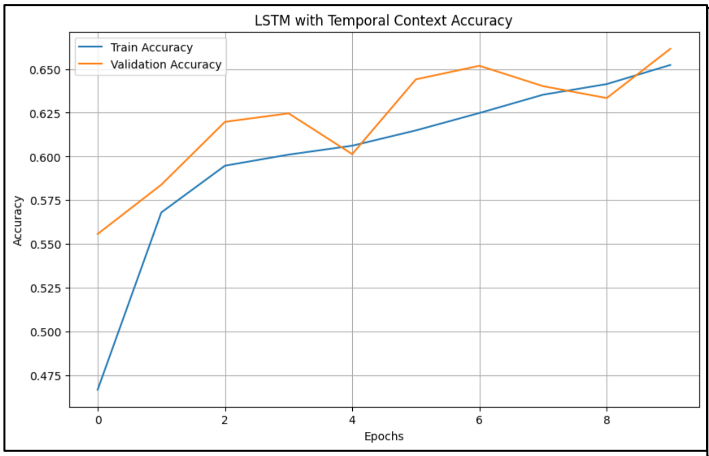
Including temporal context improved the LSTM's accuracy to **66%**.

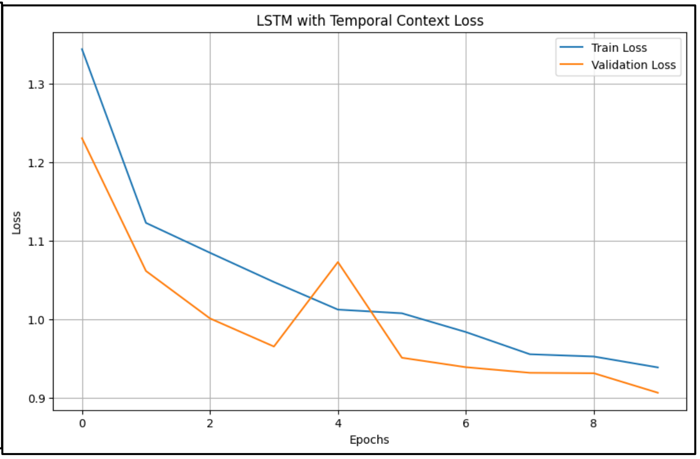
The model performed better in distinguishing **Stage 1 (N1)** and **REM (R)** by utilizing the contextual information from neighboring epochs.

- ***Training Dynamics***:

The training and validation loss curv es showed stable convergence, with no significant overfitting.

Temporal context allowed the model to better learn transitions, resulting in improved precision and recall for challenging stages.





3.3 Comparative Performance

Model	Accuracy	Key Strength	Key Weakness
RF_LOSO	68%	Simplicity and interpretability	Limited for transitional stages like N1
RF_TC	76%	Temporal context improved accuracy significantly	Computational overhead due to feature expansion
Pure LSTM	63%	Captures sequential dependencies moderately	Struggled without temporal context
LSTM_TC	66%	Better for transitions between stages	Requires large datasets for generalization

3.4 Confusion Matrix Analysis

Random Forest (RF_TC)

- **Strengths**:

  Stage 2 (N2): RF_TC achieved precision and recall above 90%, highlighting the model's ability to detect dominant stages.

  REM (R): Improved significantly with temporal context, reducing false positives into Stage 2 (N2).

- **Weaknesses**:

  Stage 1 (N1): Misclassifications into Wake (W) persisted, though reduced with TC.

LSTM_TC

- **Strengths**:

  REM (R): Leveraged temporal context to distinguish REM from Wake and N2.

  Stage 1 (N1): Transition accuracy improved, though it remained challenging.

- **Weaknesses**:

  Stage 3/4 (N3): Slightly lower recall due to its similarity to N2 in some cases.

3.5 Key Insights

1. **Temporal Context Matters**:

   Both RF and LSTM models benefited significantly from the inclusion of temporal context, with accuracy improvements of up to 8%.

   TC enabled models to better identify transitional and boundary stages like N1 and REM.

2. **Stage 2 (N2) is Dominant**:

   All models performed best on Stage 2 (N2) due to its distinct EEG features (spindles, K-complexes) and high representation in the dataset.

3. **Stage 1 (N1) Remains Challenging**:

   The transitional nature of Stage 1 and its low representation made it the hardest stage to classify accurately.

   Temporal context partially addressed this issue, but further improvements are needed.

4. **Model Trade-offs**:

   RF is simpler and interpretable but limited in handling sequential dependencies without engineered features.

   LSTM models excel at capturing temporal patterns but require larger datasets and computational resources for optimal performance.

---

4. Discussion

The study revealed several insights into sleep stage classification:

1. **Temporal Context Improves Performance**:
   - RF-TC and LSTM-TC demonstrated superior accuracy by leveraging transitions across epochs.
   - For transitional stages like Stage 1, temporal dependencies proved crucial.
2. **Stage 2 is the Easiest to Classify**:
   - Its unique EEG features (spindles, K-complexes) made it highly distinguishable, resulting in high precision and recall.
3. **Challenges in Classifying Stage 1**:
   - Stage 1 shares similarities with Wake (low alpha activity) and Stage 2 (onset of theta activity), making it hard to classify.
   - Class imbalance further contributed to the difficulty.
4. **Model Comparisons**:

- RF is interpretable and performed well on dominant stages (e.g., Stage 2), but struggled with temporal transitions.
- LSTM, particularly with TC, better handled sequential data and transitions between stages.

5. Conclusion

This study aimed to evaluate and compare the performance of traditional machine learning and deep learning methods for sleep stage classification using EEG data. The analysis focused on two primary objectives: generalizability across participants and the impact of incorporating temporal context on model performance. Using the PhysioNet Sleep-EDF dataset, the models were tested across a diverse range of participants and evaluated for their ability to classify five sleep stages accurately. The results underline the importance of generalization in sleep stage classification models. Using Leave-One-Subject-Out (LOSO) cross-validation, the models were tested on unseen participants, simulating real-world scenarios. While both RF and LSTM benefitted from temporal context, RF consistently outperformed LSTM in terms of accuracy and computational efficiency. This suggests that RF, with engineered features and temporal context, remains a competitive approach for small to medium-sized datasets.

Reference

1. Mourtazaev, M. S., Kemp, B., Zwinderman, A. H., & Kamphuisen, H. A. C. (1995). *Age and Gender Affect Different Characteristics of Slow Waves in the Sleep EEG*. PhysioNet Sleep-EDF Dataset. https://pubmed.ncbi.nlm.nih.gov/8552926/

2. Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32. https://link.springer.com/article/10.1023/A:1010933404324

3. Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830. https://dl.acm.org/doi/pdf/10.5555/1953048.2078195

4. MNE Developers. (2023). *"Overview of MNE-Python: Tutorials and Example Codes."*
Available at: https://mne.tools/dev/auto_tutorials/intro/10_overview.html
Accessed: [Insert the date you accessed the tutorial].