Classification of Sleep Stages from EEG Data

Iulia Maria GHEJU*

Computer Science Department, Faculty of Computer Science, West University of Timişoara, Bd. Vasile Pârvan no. 4, 300223 Timişoara, Romania
E-mail: iulia.gheju00@e-uvt.ro

* Author to whom correspondence should be addressed;

Abstract

Background: Manual sleep staging from electroencephalography (EEG) is time-consuming and affected by inter-scorer variability. Automated sleep stage classification can support clinical workflows if it generalizes across subjects, remains computationally feasible, and provides reliable probability estimates suitable for downstream decision-making. *Methods*: I developed a reproducible pipeline for classifying 30-second EEG epochs into W, N1, N2, N3, and REM using open sleep datasets. The pipeline standardizes preprocessing (epoching, normalization, and optional filtering), enforces subject-wise train/validation/test splits to evaluate generalization to unseen individuals, and reports accuracy, macro-F1, and Cohen's kappa alongside confusion matrices and per-class metrics. Five approaches were implemented and compared: (1) feature-based classical machine learning (logistic regression, support vector machines, random forests); (2) a 1D convolutional neural network (CNN) trained on raw epochs; (3) a context-aware CNN–BiLSTM model that combines per-epoch embeddings with bidirectional temporal modeling; (4) a sequence-to-sequence encoder–decoder with attention to jointly label consecutive epochs; and (5) Hidden Markov Model post-processing to smooth predictions and penalize physiologically implausible transitions. Reliability was assessed via probability calibration analyses. *Results*: The pipeline produces consistent, subject-wise comparable evaluations across model families and enables analysis of typical failure modes (e.g., confusion between adjacent stages). Comparative experiments quantify trade-offs between feature-based and end-to-end models, short-context and long-context temporal modeling, and the effect of HMM smoothing and calibration on temporal consistency and reliability. *Conclusions*: A structured, reproducible evaluation framework clarifies how architectural choices and post-processing influence accuracy, generalization, and reliability in EEG-based sleep staging, supporting evidence-based selection of models for medical-informatics applications.

Keywords: Automatic sleep staging; EEG; Deep Learning; Sequence modeling; Subject-wise evaluation; Medical informatics.