Sleep Stage Classification: Investigating Deep Learning with High Dataset Imbalance Under Temporal Interval Change and Data Imbalance Techniques

Abstract:

Sleep stage scoring is the process of assigning sleep stages to periods of sleep depending on features of electrophysiological signals. Automated Sleep Stage Classification (ASSC) using machine learning aims to improve efficiency of manual sleep scoring by experts, since sleep stage scoring is tedious and time-consuming when done manually. This study aims to investigate deep learning techniques through CNN-GRU architecture, primarily aiming to investigate the effects of temporal interval of data and statistical techniques to address dataset heavy imbalance. Three channel (two EEG plus EOG) was found to be superior to two channel (only two EEG) in all models tested. A 60s temporal interval data two-stage bypass model with N1 data augmentation utilizing 3 channels was found to have the best average performance with macro f1 score 0.78 ± 0.02 and macro N1 f1 score 0.51 ± 0.03, but certain 90s temporal interval models and two-stage variants showed statistically similar performance. The results reaffirm the superiority of a multimodal approach (three channel vs two channel) and the importance of context (60s and 90s vs 30s). It is evident that sleep stage scoring can be conducted efficiently and accurately through ASSC.

Keywords: Sleep, Sleep Stage Classification, Deep Learning

Introduction:

Sleep has been noted to be essential for physiological well-being [1]. It is estimated that sleep accounts for approximately ⅓ of a person's life [2], and it is evident that sleep is vital to physiological well-being during the day [3]. During sleep, the brain becomes much less active while conducting cleaning processes through physiological mechanisms, where the interstitial space between brain cells may expand up to 60%, allowing for more free flow of cerebrospinal fluid (CSF) to remove waste products such as beta-amyloid and tau proteins [4], demonstrating one way proper sleep is essential for human health. Sleep disorders are prevalent but poorly diagnosed in the general population globally [5], and sleep disorders may have the capacity to cause serious negative health outcomes; for example, trouble falling asleep may lead to neurocognitive deficits, cardiovascular problems, diabetes, recurrent heart attacks, and stroke [6] [7]. Therefore, it is imperative to continue advancing sleep medicine and research for human health.

Sleep Stages and Classification

Sleep is divided into sleep stages to identify and focus on variations in sleep. Rechtschaffen and Kales (R&K) in 1968 was the first systematic manual for classifying sleep stages based upon physiological signals, dividing sleep into six stages: Wake, Sleep Stage 1 (N1), Sleep Stage 2 (N2), Sleep Stage 3 (N3), Sleep Stage 4 (N4), and Rapid Eye Movement (REM) [8]. Older databases may use the R&K scoring manual. More recently, the American Academy of Sleep Medicine (AASM) recognizes only five sleep stages, merging N3 and N4 into a single stage, N3, termed "slow-wave sleep" or "deep sleep" [9].

Polysomnography

A polysomnogram (PSG) is the traditional method used for sleep stage classification [10]. Polysomnography is an overnight recording of an individual's sleep that includes multiple physiological signals, primarily electrophysiological [10]. PSG tests require patients to remain overnight in a controlled clinical environment while attached to measuring devices [10]. The main signal used for sleep staging is the electroencephalogram (EEG). EEG is a measure of electrical activity in the brain. EEG is performed via a non-invasive procedure in which several electrodes are attached to the head at multiple sites, and electrical activity is recorded from the electrodes to a computer [11]. The internationally recognized system for EEG electrode placement is the 10-20 system, often used in EEG and PSG studies [12]. An electrooculogram (EOG) is an electrophysiological measure of the resting electrical potential between the cornea and the retina, used to evaluate the retinal pigment epithelium (RPE) [13][14]. An electromyogram (EMG) measures muscle electrical activity through either narrow needles, or more commonly utilized now is with electrodes on the desired muscle [15]. An electrocardiogram (ECG) measures heart electrical activity through electrodes placed around the body. [16].

Sleep Stage Characteristics

The previously mentioned sleep stages are measured based on the characteristics of PSG signals. Alpha waves (8-13Hz) are associated with the wake state and are characterized by consistent high-amplitude electrical activity. Sleep stage 1 is associated with both alpha waves and theta waves (4-7Hz), starting with alpha waves and transitioning to theta waves. It has characteristics similar to those of the wake stage, but with smaller amplitudes. Theta waves continue to dominate through sleep stage 2. Sleep spindles and K-complexes characterize stage 2. Sleep spindles are bursts of high-frequency waves, and K-complexes are bursts of high-amplitude waves. Stage 3 (and N4 in the Rechtschaffen & Kales) is associated with delta waves and is termed "deep sleep". Delta waves (less than 3Hz) are present in deep sleep, and other physiological processes of the individual, such as heart rate and respiration, slow down during deep sleep. Finally, REM sleep is observed at the end of a cycle. The waves here are similar to the wake state, and eye movement in this stage is rapid, while the rest of the body does not move significantly. Dreaming is observed only in REM sleep, also confirming the fact of great brain activity during sleep hours. Moreover, sleep is a continuous and dynamic process, so therefore contextual information is important between consecutive epochs (termed "forming sequences"): for example, if N3 were surrounded by adjacent N2 stages, scorers would still classify them as N2 [17], [18], [19].

Machine Learning in Sleep Stage Classification

Automatic sleep stage classification (ASSC) through machine learning methods has been proposed to improve the efficiency of sleep stage scoring [7]. ASSC enables more efficient expert sleep-stage classification, aiding in diagnosing sleep disorders and monitoring sleep [7]. Traditional ML techniques have been used in prior research, including Support Vector Machines (SVMs), random forest classifiers, K-nearest neighbors (KNN), decision trees (DTs), and Linear Discriminant Analysis (LDA) [20]. Many studies use only one EEG channel. Alickovic et al. investigates automatic sleep-stage scoring using SVM with a single EEG channel, reporting preprocessing methods, model architecture, and performance. To improve classic SVM architecture, they employed multiscale principal component analysis and the discrete wavelet transform (DWT) to extract features from EEG data, and then fed the extracted features into a

rotational SVM. Among all subjects, the sensitivity/recall was 84.48%, and the accuracy 91.1%, and Cohen's kappa coefficient was 0.88 for the five-stage sleep classification [21]. Proper and extensive preprocessing of PSG data is imperative to obtain good performance with an ML model, as the study demonstrates. Preprocessing requires different methods regarding the unique multivariate time-series nature of PSG signals. Feature extraction is commonly implemented after. Some general EEG signal feature extraction techniques include time-domain analysis [22], frequency-domain analysis [23], and time-frequency-domain analysis [24]. The reasoning to implement this step is to generate low-dimensional or the most important features for a classification model to be trained on.

Since the growth of deep learning (DL) in the 2010s, driven by increased computing power, DL techniques have been increasingly applied to sleep stage classification. DL models have been investigated to improve on earlier models using traditional ML techniques, "including algorithms like the Multi-Layers Perceptron (MLP) and its modern refinements, including Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTM), Gated Recurrent Unit (GRU), Bi-directional LSTM, Convolutional Neural Networks (CNN)" (Sekkel et al.). Many deep learning papers utilize a base model of a feature extraction step and a temporal resolution step, through a CNN and an RNN respectively [25].

Aim and Objectives

While traditional ML techniques have been used, deep learning techniques have only been implemented in relatively recent years. This study will investigate a base model deep learning architecture, with feature extraction and temporal resolution through a CNN and RNN, respectively. Additionally, imbalance techniques aimed at optimizing model performance remain a focus of research. These parts of investigating machine learning in sleep stage classification (feature extraction, context, imbalance techniques) have been thoroughly investigated separately, but this study presents a holistic analysis on these topics.

Methods:

Dataset

One of the largest and most widely used PSG datasets for sleep-stage classification is the Sleep-EDF dataset from PhysioNet [25]. It includes 197 whole-night PSG recordings of healthy Caucasian individuals aged 25-101. The database includes 2 EEG channels (Fpz-Cz & Pz-Oz electrode locations), EOG (horizontal), and submental chin EMG signals, along with oro-nasal respiratory and rectal body temperature measurements at some recording sites [25]. The Fpz-Oz EEG, Pz-Oz EEG, and EOG channels (2 EEG channels + EOG) were the only signals used in this study. The EEG and EOG were sampled at 100Hz. Each subject had 2 nights of sleep recordings. Some subjects had missing recordings for night 1 or night 2. Only night 1 sleep recordings from the first 25 subjects were used for this study. Each PSG was accompanied by a hypnogram, manually scored by expert technicians according to the R&K manual. The data were derived from two studies, a sleep cassette and a sleep telemetry study. The sleep cassette study focused on healthy subjects, whereas the sleep telemetry study focused on subjects receiving sleep medication for sleep disorders [25]. The data used in this study were obtained from the Sleep Cassette Study to focus on sleep-stage classification in healthy subjects. The sleep stages N3 and N4 from the Rechtschaffen & Kales manual were merged into a single stage, N3, for this study, following the AASM sleep stage guide.

Preprocessing

Data processing involved passing EEG and EOG signals through band-pass filters (EEG band-pass 0.3-35 Hz, EOG band-pass 0.1-15 Hz). Then, subject-wise normalization was computed for each recording using robust median/median absolute deviation. Each of the 30s epochs was then matched with the corresponding hypnogram annotations from the dataset.

The data split of the dataset was 80:10:10 for all models, for training, test, and validation data, respectively. To maintain comparability, all models of the same temporal interval used the same unified dataset, and all datasets were derived from the same exact preprocessed data, ensuring statistical rigor and comparability between models. A global random seed was set to confirm deterministic nature and statistical comparability.

Model Architecture

The model employed a convolutional neural network (CNN) [25] and a Gated Recurrent Unit (GRU) [26][27]. In summary, the first stage of the model used a CNN to extract local temporal features from signals, and a GRU was then applied to capture longer-term features and dependencies, as well as sequential patterns. Finally, the GRU embeddings were fed into a classification layer to make predictions.

To test effects of temporal window duration on model performance, different methods of data feeding were implemented. “30 seconds” refers to the temporal interval using the time division of each segment from the original database, with data feeding being one to one. “60 seconds” refers to utilizing the 30s segment prior to the current 30s segment as context during data feeding. “90 seconds”, in turn, refers to utilizing the two 30s segments prior to the current 30s segment.

Convolutional Feature Extraction

CNNs have been successfully applied to feature extraction for individual data segments [28]. A three-block one-dimensional convolutional neural network (CNN) is used to extract hierarchical temporal features from the EEG and EOG signals. This is done by downsampling time dimension while increasing feature depth by using 1D convolutions; then normalization, GELU activation, max pooling, and dropout. All convolutions use symmetric padding to preserve temporal alignment prior to pooling operations. The CNN model contains the following blocks:

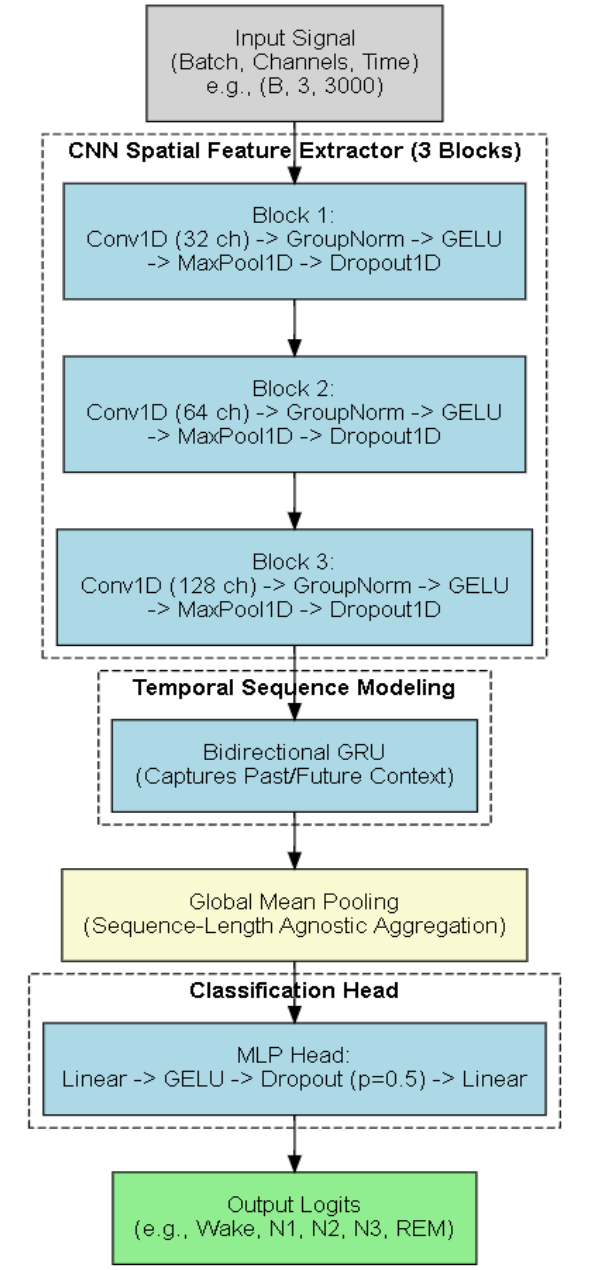

Figure 1. CNN-GRU Model Architecture

1. Spatial Feature Extraction (1D-CNN Encoder)

The CNN module is composed of three cascaded convolutional blocks designed to extract hierarchical features. It progressively reduces the temporal resolution of the signal while increasing the depth of the feature maps.

- Block 1 (Low-Frequency/Broad Feature Extraction) uses a large kernel size (kernel_size=9, stride=2, padding=4) to capture broad, low-frequency waveforms (e.g., Delta waves). It expands the input channels to 32 feature maps.
- Block 2 (Mid-Frequency Feature Extraction) uses a moderate kernel size (kernel_size=7, stride=1, padding=3) mapping 32 channels to 64 feature maps.

- Block 3 (High-Frequency/Morphological Feature Extraction) uses a smaller kernel size (kernel_size=5, stride=1, padding=2) mapping 64 channels to 128 feature maps to capture fast, transient micro-events like sleep spindles or K-complexes.

Regularization and Activation within CNN Blocks: Each convolutional operation is immediately followed by:

1. Normalization: A dynamically selected normalization layer (defaulting to GroupNorm with 8 groups, or standard BatchNorm1d). GroupNorm is particularly advantageous for biological time-series where batch sizes might be constrained.
2. Activation: Gaussian Error Linear Unit (GELU) avoids vanishing gradient problem associated with hard-zero activations like ReLU, providing smoother optimization landscape.
3. Downsampling: A MaxPool1d(kernel_size=2) layer halves temporal dimension, distilling the most salient features and significantly reducing computational complexity for the RNN.
4. Spatial Dropout: Dropout1d(p=0.25) layer drops entire feature maps randomly, enforcing robust feature independence across the spatial filters.

2. Temporal Sequence Modeling (Bi-GRU)

Following the CNN, the feature tensor is transposed from (Batch, Features, Time) to (Batch, Time, Features). It is then fed into a Bidirectional Gated Recurrent Unit (Bi-GRU):

- Architecture: Bi-GRU consists of a default hidden size of 96 (rnn_hidden=96).
- Bidirectionality: By processing the sequence both forwards and backwards, the network captures contextual dependencies from both past and future morphological events within the epoch.
- Output: Bi-GRU concatenates the forward and backward hidden states, resulting in a dense temporal feature representation of size 2 * rnn_hidden = (192) at every downsampled timestep.

3. Global Temporal Pooling

To transition from a sequence of temporal states to a single classification decision, Global Mean Pooling (z.mean(dim=1)) was implemented to take the average across all timesteps, producing a highly stable, sequence-length-agnostic representation of the entire epoch. This architecture could be extended to test longer temporal intervals.

$$z = \frac{1}{T'} \sum_{t=1}^{T'} h_t \quad \text{(Eqn 1)}$$

4. Classification Head (MLP)

The globally pooled feature vector is then passed through a Multi-Layer Perceptron (MLP) classification head:

1. Post-RNN Dropout: A heavy dropout layer (p=0.5) is applied to the pooled representation to prevent overfitting on specific temporal patterns.
2. Dense Layer 1: Projects the 192-dimensional vector to 128 dimensions, followed by a GELU activation and another 50% Dropout.

3. Dense Layer 2 (Output): A final linear layer maps the 128-dimensional vector to the required n_classes logits (e.g., 5 for standard staging, 2 for the binary Wake/Sleep stage).

Imbalance Techniques:

Some models in this study utilize techniques that attempt to address the problem of dataset imbalance, here the Wake stage (and N2 to a lesser extent) to N1 (and N3 to a lesser extent). The relationship between wake, N1, and N2 will be the target of the stated methods. The techniques to attempt to rectify the dataset imbalance include the Weighted Random Sampler class from PyTorch (referenced here as "WeightedRandomSampler" or "Sampler"), weighted cross entropy loss (referenced here as "weighted loss"), focal loss, and N1 data augmentation. Weighted Random Sampler oversamples the minority class, in this case the N1 stage, aiming to address the imbalance through the data loading stage rather than during loss calculation. Weighted loss and focal loss, on the other hand, attempt to address the imbalance through penalizing misclassifications of minority classes more severely than misclassification of majority classes. In other words, the model cannot reach a local minimum easily by guessing the majority class. This method is used more for higher data imbalances, which would lend itself useful here. Data augmentation physically increases the volume and variety of the minority class during training. Augmented data will be slight variations of the minority class, addressing the dataset imbalance through directing the model to learn slight fluctuations in data from creation of slightly varied data.

Study Design

First, model performance on three-channel (two EEG + one EOG) and two-channel (only two EEG) data was compared to confirm the superiority of a multimodal or multiple channel approach, and will be referred to as "two channel" or "three channel". Then, a comparison between a single five class classifier versus a bypass system with two models, which will be referred to as "single" models and "two stage" models respectively. The "single" type is straightforwardly a single classifier of the five sleep stages. The "two stage" model utilizes a bypass system, with "stage one" being a binary classifier between wake and non-wake. "Stage two" would then be a four class non-wake classifier, taking in data classified as non-wake by the first binary classifier. Each "stage" would have the corresponding model labeling, i.e. binary classifier having wake versus non-wake and the four class non-wake classifier having every specific sleep stage excluding wake. This model design was implemented to test whether such an architecture could statistically significantly improve model performance. Finally, the aforementioned temporal intervals 30s, 60s, and 90s would be implemented to test the effect of context. At the end of this process, there would be 9 core models, a two channel single, three channel single, and three channel two stage model for every tested temporal interval. To further test possible model improvements, the four aforementioned imbalance techniques would be tested for every temporal interval and for single and two stage, producing eight models for every temporal interval. Lastly, a three channel single and two stage downsampled model following the "30 minute buffer rule" or rule of removing all wake samples from a dataset other than the first and last 30 minutes of a PSG was trained and tested to compare the viability of high class imbalance utilized in the notebook.

Results:

Statistical analysis was carried out through subject-level bootstrapping and 5-fold cross validation, which are shown through the below tables and graphs. Subject-level bootstrapping and K-fold cross validation were used (instead of epoch-level) to account for subject variance and to limit the possibility of model overfitting.

The downsampled model was included as a comparison to methodology in contemporary studies, following the "30 minute buffer rule", or only including the initial and ending 30 minutes of wake stage in a PSG recording to lessen the wake/non-wake imbalance.

Table 1. 30s 2ch Single, 30s 3ch Single, 30s 3ch 2-Stage Models

Model	30s 2Ch Single	30s 3Ch Single	30s 3Ch 2-Stage
Accuracy	0.9065 ± 0.0369	0.9135 ± 0.0331	0.9170 ± 0.0308
Macro Precision	0.7100 ± 0.0752	0.7244 ± 0.0853	0.7639 ± 0.0417
Macro Recall	0.6754 ± 0.0655	0.6956 ± 0.0639	0.7295 ± 0.0279
Macro F1	0.6771 ± 0.0718	0.6898 ± 0.0796	0.7359 ± 0.0259
Wake Precision	0.9617 ± 0.0341	0.9661 ± 0.0265	0.9771 ± 0.0254
Wake Recall	0.9934 ± 0.0031	0.9916 ± 0.0025	0.9867 ± 0.0052
Wake F1	0.9770 ± 0.0171	0.9784 ± 0.0129	0.9817 ± 0.0121
N1 Precision	0.3067 ± 0.1783	0.3770 ± 0.2144	0.4730 ± 0.1063
N1 Recall	0.1787 ± 0.1103	0.1726 ± 0.1904	0.3824 ± 0.0776
N1 F1	0.2238 ± 0.1334	0.2108 ± 0.1998	0.4125 ± 0.0568
N2 Precision	0.8140 ± 0.0870	0.8363 ± 0.0431	0.8328 ± 0.0444
N2 Recall	0.8578 ± 0.0868	0.8457 ± 0.0513	0.8720 ± 0.0557
N2 F1	0.8276 ± 0.0467	0.8394 ± 0.0309	0.8494 ± 0.0220
N3 Precision	0.7719 ± 0.2165	0.7526 ± 0.2204	0.8142 ± 0.1541
N3 Recall	0.7395 ± 0.1036	0.7730 ± 0.0441	0.6522 ± 0.1188
N3 F1	0.7229 ± 0.1080	0.7376 ± 0.1428	0.7024 ± 0.0757
REM Precision	0.6953 ± 0.1126	0.6900 ± 0.1192	0.7223 ± 0.0796
REM Recall	0.6076 ± 0.2148	0.6953 ± 0.1573	0.7540 ± 0.0741
REM F1	0.6340 ± 0.1778	0.6827 ± 0.1249	0.7334 ± 0.0490

Table 2. 60s 2ch Single, 60s 3ch Single, 60s 3ch 2-Stage Models

Model	60s 2Ch Single	60s 3Ch Single	60s 3Ch 2-Stage
Accuracy	0.9311 ± 0.0055	0.9371 ± 0.0055	0.9363 ± 0.0048
Macro Precision	0.7743 ± 0.0166	0.7891 ± 0.0176	0.7887 ± 0.0110
Macro Recall	0.7077 ± 0.0110	0.7513 ± 0.0127	0.7505 ± 0.0116
Macro F1	0.7125 ± 0.0105	0.7618 ± 0.0128	0.7599 ± 0.0105
Wake Precision	0.9870 ± 0.0027	0.9877 ± 0.0028	0.9905 ± 0.0024
Wake Recall	0.9919 ± 0.0023	0.9915 ± 0.0022	0.9865 ± 0.0031
Wake F1	0.9894 ± 0.0015	0.9896 ± 0.0016	0.9885 ± 0.0017
N1 Precision	0.5153 ± 0.0631	0.5124 ± 0.0684	0.5263 ± 0.0399
N1 Recall	0.1034 ± 0.0141	0.2733 ± 0.0395	0.2662 ± 0.0355
N1 F1	0.1714 ± 0.0200	0.3538 ± 0.0395	0.3517 ± 0.0328
N2 Precision	0.8392 ± 0.0217	0.8538 ± 0.0224	0.8506 ± 0.0196
N2 Recall	0.8936 ± 0.0193	0.8930 ± 0.0177	0.9069 ± 0.0130
N2 F1	0.8652 ± 0.0130	0.8727 ± 0.0129	0.8777 ± 0.0112
N3 Precision	0.8414 ± 0.0432	0.8238 ± 0.0455	0.8374 ± 0.0425
N3 Recall	0.8108 ± 0.0413	0.8376 ± 0.0344	0.8024 ± 0.0416
N3 F1	0.8246 ± 0.0285	0.8295 ± 0.0279	0.8184 ± 0.0289
REM Precision	0.6887 ± 0.0328	0.7677 ± 0.0237	0.7389 ± 0.0253
REM Recall	0.7390 ± 0.0378	0.7610 ± 0.0339	0.7906 ± 0.0219
REM F1	0.7118 ± 0.0204	0.7636 ± 0.0180	0.7633 ± 0.0121

Table 3. 90s 2ch Single, 90s 3ch Single, 90s 3ch 2-Stage Models

Model	90s 2Ch Single	90s 3Ch Single	90s 3Ch 2-Stage
Accuracy	0.9249 ± 0.0116	0.9286 ± 0.0098	0.9290 ± 0.0098
Macro Precision	0.7658 ± 0.0190	0.7945 ± 0.0129	0.7841 ± 0.0122
Macro Recall	0.7285 ± 0.0137	0.7429 ± 0.0138	0.7610 ± 0.0149
Macro F1	0.7400 ± 0.0137	0.7617 ± 0.0125	0.7692 ± 0.0124
Wake Precision	0.9807 ± 0.0054	0.9836 ± 0.0040	0.9900 ± 0.0038
Wake Recall	0.9907 ± 0.0023	0.9907 ± 0.0022	0.9856 ± 0.0029
Wake F1	0.9857 ± 0.0032	0.9871 ± 0.0025	0.9878 ± 0.0026
N1 Precision	0.4413 ± 0.0531	0.5421 ± 0.0372	0.5042 ± 0.0353
N1 Recall	0.2406 ± 0.0167	0.3267 ± 0.0380	0.3653 ± 0.0457
N1 F1	0.3101 ± 0.0196	0.4059 ± 0.0309	0.4216 ± 0.0342
N2 Precision	0.8470 ± 0.0173	0.8257 ± 0.0209	0.8277 ± 0.0207
N2 Recall	0.8669 ± 0.0248	0.8863 ± 0.0172	0.8848 ± 0.0166
N2 F1	0.8566 ± 0.0154	0.8547 ± 0.0125	0.8551 ± 0.0128
N3 Precision	0.8187 ± 0.0434	0.8177 ± 0.0492	0.8131 ± 0.0441
N3 Recall	0.7440 ± 0.0430	0.7216 ± 0.0447	0.7624 ± 0.0347
N3 F1	0.7782 ± 0.0295	0.7650 ± 0.0304	0.7856 ± 0.0240
REM Precision	0.7412 ± 0.0346	0.8035 ± 0.0351	0.7857 ± 0.0335
REM Recall	0.8002 ± 0.0302	0.7894 ± 0.0254	0.8070 ± 0.0217
REM F1	0.7692 ± 0.0285	0.7957 ± 0.0203	0.7957 ± 0.0192

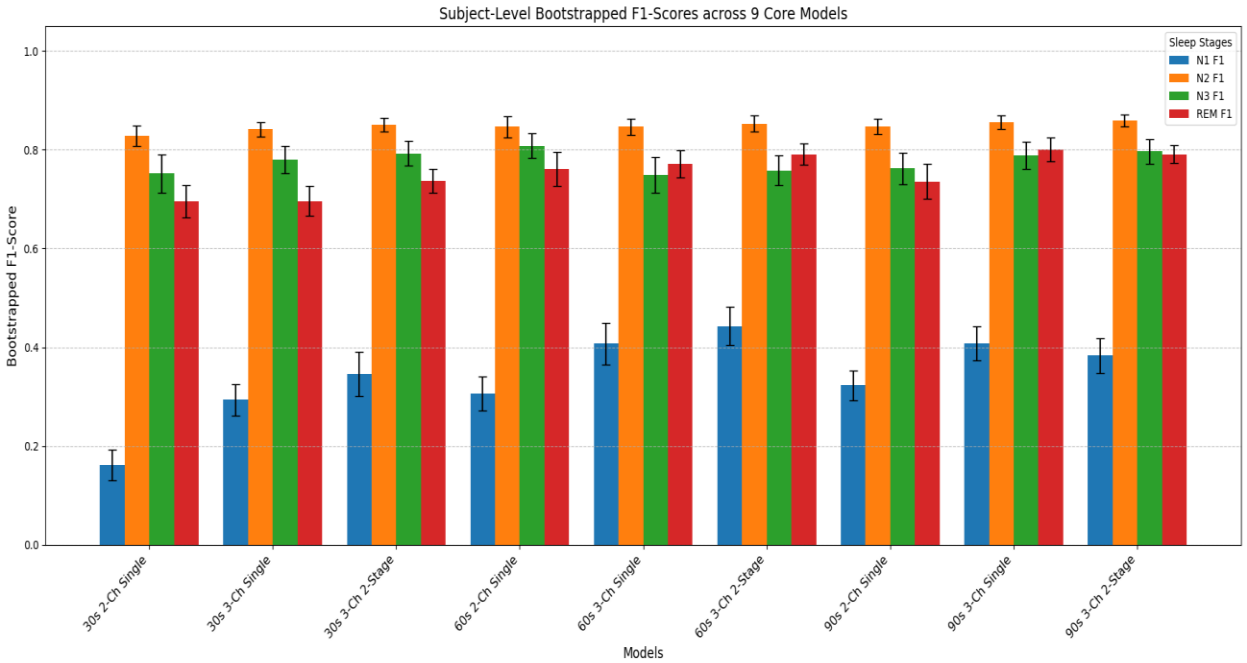

Figure 2. Bootstrapping Across 9 Core Models

Table 4. 30s 3ch Single Imbalance Technique Models

Model	30s Single Sampler	30s Single Weighted	30s Single Focal	30s Single AugN1
Accuracy Mean	0.902886	0.89075	0.888278	0.921568
Accuracy Std	0.012895	0.012753	0.015329	0.013425
Accuracy 95% CI	[0.8768, 0.9263]	[0.8638, 0.9151]	[0.8580, 0.9167]	[0.8918, 0.9450]
Macro Precision Mean	0.726816	0.696064	0.70654	0.772997
Macro Precision Std	0.013187	0.013935	0.015935	0.013906
Macro Precision 95% CI	[0.7002, 0.7522]	[0.6678, 0.7229]	[0.6743, 0.7363]	[0.7443, 0.7985]
Macro Recall Mean	0.814574	0.777801	0.798449	0.755679
Macro Recall Std	0.011403	0.013982	0.013105	0.020833
Macro Recall 95% CI	[0.7911, 0.8361]	[0.7509, 0.8049]	[0.7702, 0.8227]	[0.7099, 0.7916]
Macro F1 Mean	0.751104	0.71977	0.723383	0.763197
Macro F1 Std	0.013319	0.01328	0.016973	0.01592

Macro F1 95% CI	[0.7237, 0.7771]	[0.6931, 0.7440]	[0.6881, 0.7544]	[0.7290, 0.7913]
Wake Precision Mean	0.995682	0.996329	0.992378	0.976855
Wake Precision Std	0.001905	0.001136	0.004823	0.013523
Wake Precision 95% CI	[0.9914, 0.9986]	[0.9937, 0.9982]	[0.9814, 0.9983]	[0.9463, 0.9931]
Wake Recall Mean	0.962027	0.955225	0.956362	0.989204
Wake Recall Std	0.006535	0.007147	0.009653	0.00218
Wake Recall 95% CI	[0.9479, 0.9731]	[0.9400, 0.9676]	[0.9354, 0.9727]	[0.9847, 0.9930]
Wake F1 Mean	0.978554	0.975331	0.974009	0.982944
Wake F1 Std	0.003635	0.003903	0.005802	0.007082
Wake F1 95% CI	[0.9707, 0.9849]	[0.9670, 0.9822]	[0.9609, 0.9835]	[0.9666, 0.9922]
N1 Precision Mean	0.304743	0.269601	0.279845	0.436701
N1 Precision Std	0.026421	0.02719	0.025883	0.037321
N1 Precision 95% CI	[0.2531, 0.3557]	[0.2138, 0.3213]	[0.2278, 0.3294]	[0.3576, 0.5059]
N1 Recall Mean	0.702967	0.609222	0.72023	0.436311
N1 Recall Std	0.03046	0.039325	0.032867	0.041548
N1 Recall 95% CI	[0.6408, 0.7575]	[0.5263, 0.6806]	[0.6521, 0.7799]	[0.3564, 0.5145]
N1 F1 Mean	0.424446	0.373004	0.40234	0.435194
N1 F1 Std	0.027305	0.029574	0.02837	0.031505
N1 F1 95% CI	[0.3704, 0.4770]	[0.3105, 0.4273]	[0.3436, 0.4531]	[0.3722, 0.4942]
N2 Precision Mean	0.902402	0.865082	0.898077	0.848503
N2 Precision Std	0.014763	0.020177	0.016223	0.019712
N2 Precision 95% CI	[0.8749, 0.9307]	[0.8277, 0.9044]	[0.8664, 0.9282]	[0.8094, 0.8854]
N2 Recall Mean	0.75364	0.758365	0.712334	0.842067
N2 Recall Std	0.033269	0.027197	0.036801	0.032171
N2 Recall 95% CI	[0.6838, 0.8113]	[0.7005, 0.8051]	[0.6351, 0.7776]	[0.7754, 0.8945]
N2 F1 Mean	0.820866	0.807766	0.793841	0.844789

N2 F1 Std	0.020576	0.015993	0.023053	0.017816
N2 F1 95% CI	[0.7746, 0.8548]	[0.7742, 0.8371]	[0.7416, 0.8327]	[0.8056, 0.8760]
N3 Precision Mean	0.719544	0.669749	0.596597	0.814023
N3 Precision Std	0.051238	0.055412	0.067984	0.04683
N3 Precision 95% CI	[0.6137, 0.8086]	[0.5516, 0.7684]	[0.4571, 0.7156]	[0.7114, 0.8926]
N3 Recall Mean	0.874665	0.859876	0.903091	0.781106
N3 Recall Std	0.022212	0.033186	0.022688	0.035541
N3 Recall 95% CI	[0.8334, 0.9194]	[0.7930, 0.9216]	[0.8556, 0.9445]	[0.7096, 0.8498]
N3 F1 Mean	0.788187	0.751149	0.716029	0.795855
N3 F1 Std	0.029395	0.034705	0.050084	0.025337
N3 F1 95% CI	[0.7237, 0.8380]	[0.6754, 0.8087]	[0.6049, 0.7985]	[0.7408, 0.8399]
REM Precision Mean	0.711708	0.679561	0.765803	0.788904
REM Precision Std	0.039189	0.042141	0.032366	0.03636
REM Precision 95% CI	[0.6304, 0.7817]	[0.5864, 0.7544]	[0.6960, 0.8227]	[0.7144, 0.8492]
REM Recall Mean	0.779571	0.706315	0.700227	0.729707
REM Recall Std	0.029375	0.032046	0.035972	0.036436
REM Recall 95% CI	[0.7214, 0.8353]	[0.6424, 0.7673]	[0.6254, 0.7703]	[0.6509, 0.7936]
REM F1 Mean	0.743464	0.6916	0.730694	0.757202
REM F1 Std	0.028061	0.026428	0.024087	0.025179
REM F1 95% CI	[0.6816, 0.7920]	[0.6334, 0.7376]	[0.6798, 0.7744]	[0.6989, 0.8014]

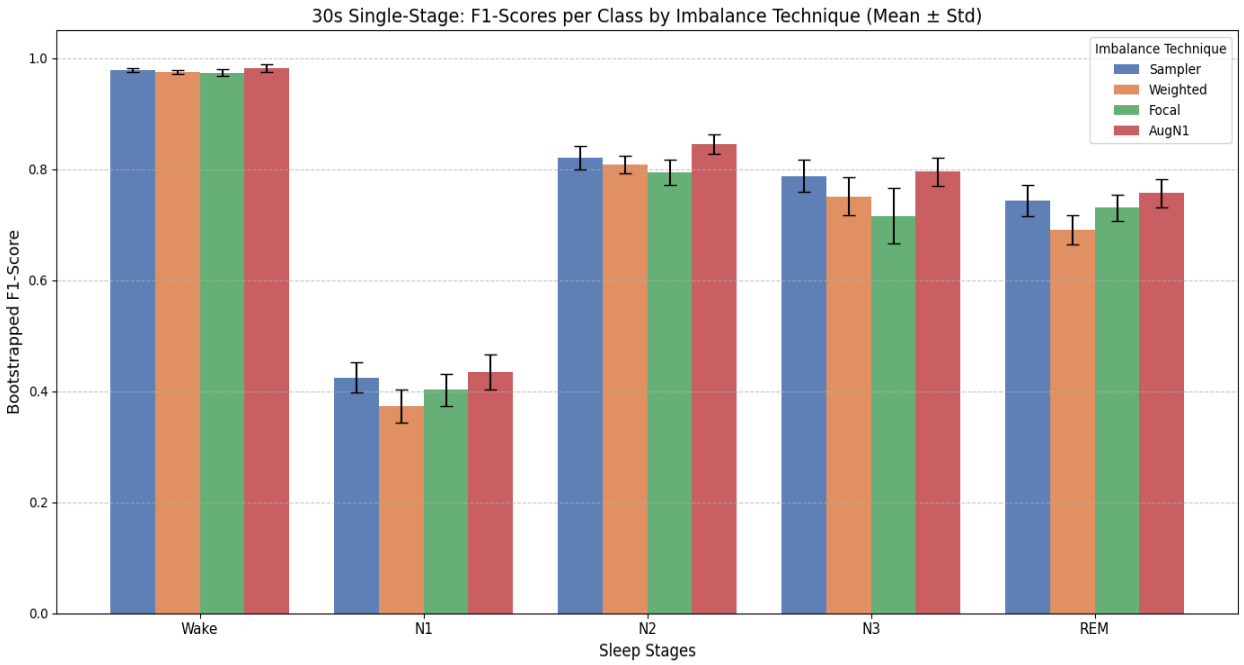

Figure 3. 30s 3ch Single Imbalance Technique Models Bootstrapping

Table 5. 30s 3ch 2-stage Imbalance Technique Models

Model	30s 2-Stage Sampler	30s 2-Stage Weighted	30s 2-Stage Focal	30s 2-Stage AugN1
Accuracy Mean	0.908556	0.90498	0.901162	0.914292
Accuracy Std	0.01434	0.01388	0.014688	0.013539
Accuracy 95% CI	[0.8795, 0.9332]	[0.8770, 0.9300]	[0.8706, 0.9276]	[0.8866, 0.9386]
Macro Precision Mean	0.729236	0.717046	0.710076	0.750172
Macro Precision Std	0.013904	0.014087	0.013379	0.013512
Macro Precision 95% CI	[0.6995, 0.7549]	[0.6883, 0.7432]	[0.6819, 0.7345]	[0.7217, 0.7742]
Macro Recall Mean	0.763853	0.738615	0.747476	0.743298
Macro Recall Std	0.021059	0.018607	0.020926	0.02104
Macro Recall 95% CI	[0.7185, 0.7975]	[0.6985, 0.7696]	[0.7029, 0.7830]	[0.6996, 0.7799]
Macro F1 Mean	0.740246	0.720941	0.718486	0.74414

Macro F1 Std	0.016494	0.015107	0.015821	0.015694
Macro F1 95% CI	[0.7027, 0.7695]	[0.6874, 0.7494]	[0.6843, 0.7472]	[0.7113, 0.7717]
Wake Precision Mean	0.975353	0.975353	0.975353	0.975353
Wake Precision Std	0.011899	0.011899	0.011899	0.011899
Wake Precision 95% CI	[0.9484, 0.9919]	[0.9484, 0.9919]	[0.9484, 0.9919]	[0.9484, 0.9919]
Wake Recall Mean	0.987701	0.987701	0.987701	0.987701
Wake Recall Std	0.002378	0.002378	0.002378	0.002378
Wake Recall 95% CI	[0.9826, 0.9920]	[0.9826, 0.9920]	[0.9826, 0.9920]	[0.9826, 0.9920]
Wake F1 Mean	0.981451	0.981451	0.981451	0.981451
Wake F1 Std	0.006218	0.006218	0.006218	0.006218
Wake F1 95% CI	[0.9672, 0.9906]	[0.9672, 0.9906]	[0.9672, 0.9906]	[0.9672, 0.9906]
N1 Precision Mean	0.354254	0.313239	0.300518	0.356597
N1 Precision Std	0.030237	0.030718	0.030171	0.028831
N1 Precision 95% CI	[0.2944, 0.4124]	[0.2533, 0.3739]	[0.2407, 0.3582]	[0.3029, 0.4159]
N1 Recall Mean	0.463533	0.417993	0.448271	0.446455
N1 Recall Std	0.054482	0.0553	0.056612	0.058297
N1 Recall 95% CI	[0.3522, 0.5596]	[0.3056, 0.5176]	[0.3336, 0.5524]	[0.3335, 0.5483]
N1 F1 Mean	0.399536	0.355759	0.357863	0.394138
N1 F1 Std	0.029022	0.029337	0.030617	0.029526
N1 F1 95% CI	[0.3382, 0.4527]	[0.2944, 0.4111]	[0.2964, 0.4151]	[0.3337, 0.4479]
N2 Precision Mean	0.90428	0.868744	0.897728	0.860902
N2 Precision Std	0.015589	0.017981	0.016505	0.019262
N2 Precision 95% CI	[0.8742, 0.9338]	[0.8316, 0.9017]	[0.8665, 0.9282]	[0.8239, 0.8978]
N2 Recall Mean	0.73047	0.765044	0.711449	0.824943

N2 Recall Std	0.033007	0.026804	0.031038	0.031755
N2 Recall 95% CI	[0.6581, 0.7858]	[0.7088, 0.8126]	[0.6481, 0.7655]	[0.7566, 0.8763]
N2 F1 Mean	0.807596	0.813236	0.793312	0.842052
N2 F1 Std	0.020111	0.016455	0.019457	0.017663
N2 F1 95% CI	[0.7633, 0.8395]	[0.7790, 0.8418]	[0.7508, 0.8260]	[0.8022, 0.8707]
N3 Precision Mean	0.713366	0.692	0.651964	0.812941
N3 Precision Std	0.049165	0.051049	0.05369	0.042574
N3 Precision 95% CI	[0.6066, 0.7968]	[0.5825, 0.7818]	[0.5438, 0.7466]	[0.7215, 0.8856]
N3 Recall Mean	0.884528	0.902271	0.906977	0.77686
N3 Recall Std	0.027555	0.018821	0.022719	0.041703
N3 Recall 95% CI	[0.8247, 0.9336]	[0.8638, 0.9373]	[0.8566, 0.9471]	[0.6909, 0.8580]
N3 F1 Mean	0.788525	0.782019	0.757112	0.793108
N3 F1 Std	0.030885	0.032593	0.036396	0.026515
N3 F1 95% CI	[0.7185, 0.8403]	[0.7075, 0.8369]	[0.6794, 0.8210]	[0.7355, 0.8396]
REM Precision Mean	0.698925	0.735894	0.72482	0.745068
REM Precision Std	0.041096	0.038158	0.034182	0.049342
REM Precision 95% CI	[0.6133, 0.7698]	[0.6545, 0.8015]	[0.6553, 0.7870]	[0.6419, 0.8271]
REM Recall Mean	0.753033	0.620068	0.682984	0.680532
REM Recall Std	0.035871	0.028674	0.036989	0.036727
REM Recall 95% CI	[0.6835, 0.8222]	[0.5644, 0.6752]	[0.6073, 0.7555]	[0.6099, 0.7539]
REM F1 Mean	0.724119	0.672241	0.70269	0.709951
REM F1 Std	0.030241	0.023709	0.02946	0.029488
REM F1 95% CI	[0.6604, 0.7788]	[0.6241, 0.7157]	[0.6409, 0.7562]	[0.6489, 0.7658]

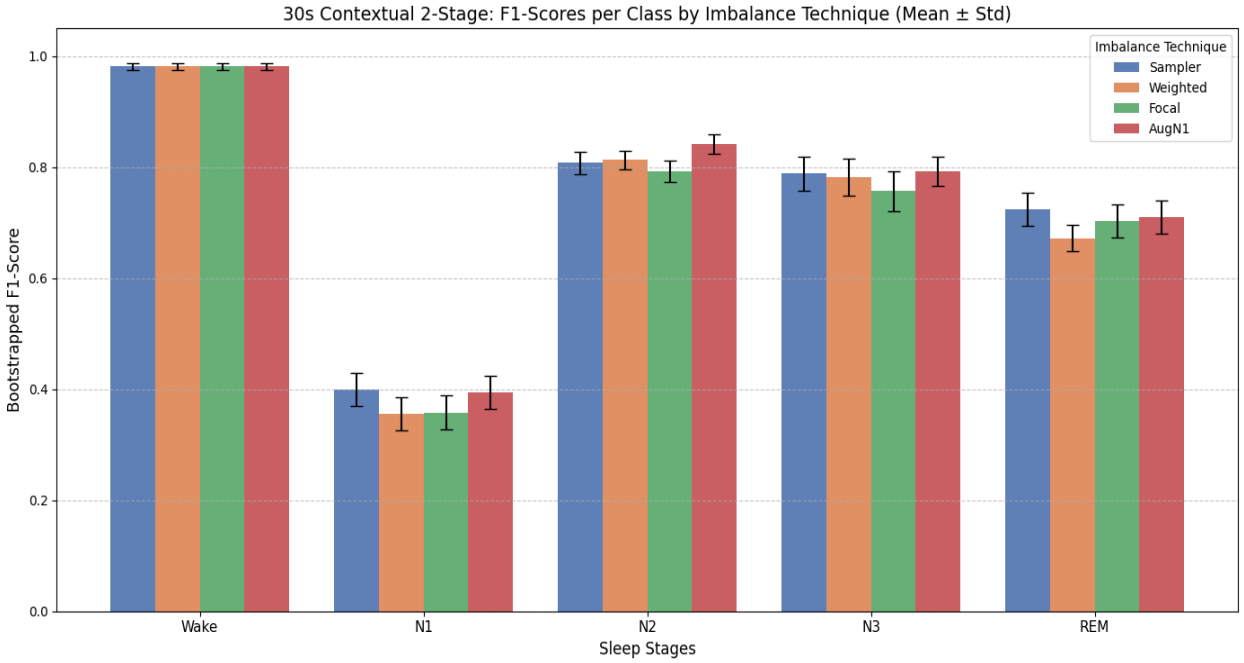

Figure 4. 30s 3ch 2-stage Imbalance Technique Models

Table 5. 60s 3ch Single Imbalance Technique Models

Model	60s Single Sampler	60s Single Weighted	60s Single Focal	60s Single AugN1
Accuracy Mean	0.903507	0.881152	0.897644	0.925624
Accuracy Std	0.012004	0.014642	0.013574	0.013697
Accuracy 95% CI	[0.8805, 0.9271]	[0.8514, 0.9097]	[0.8711, 0.9234]	[0.8950, 0.9492]
Macro Precision Mean	0.726566	0.688134	0.722546	0.796015
Macro Precision Std	0.014348	0.016855	0.014285	0.015013
Macro Precision 95% CI	[0.6958, 0.7526]	[0.6554, 0.7187]	[0.6928, 0.7485]	[0.7648, 0.8240]
Macro Recall Mean	0.829468	0.775171	0.823269	0.767627
Macro Recall Std	0.012691	0.011805	0.012427	0.024132
Macro Recall 95% CI	[0.8036, 0.8539]	[0.7510, 0.7991]	[0.7978, 0.8464]	[0.7152, 0.8088]
Macro F1 Mean	0.755035	0.706433	0.745067	0.778819
Macro F1 Std	0.013576	0.016742	0.014859	0.018738

Macro F1 95% CI	[0.7264, 0.7796]	[0.6711, 0.7367]	[0.7142, 0.7723]	[0.7390, 0.8117]
Wake Precision Mean	0.995879	0.994501	0.996121	0.980071
Wake Precision Std	0.002296	0.002645	0.002264	0.012824
Wake Precision 95% CI	[0.9908, 0.9993]	[0.9885, 0.9986]	[0.9907, 0.9991]	[0.9511, 0.9951]
Wake Recall Mean	0.959777	0.946797	0.960219	0.988155
Wake Recall Std	0.007563	0.01058	0.007397	0.002448
Wake Recall 95% CI	[0.9437, 0.9729]	[0.9240, 0.9649]	[0.9448, 0.9732]	[0.9832, 0.9926]
Wake F1 Mean	0.977479	0.970033	0.977827	0.984054
Wake F1 Std	0.004204	0.005883	0.004238	0.006721
Wake F1 95% CI	[0.9686, 0.9848]	[0.9576, 0.9801]	[0.9689, 0.9851]	[0.9684, 0.9929]
N1 Precision Mean	0.320056	0.260851	0.309408	0.494889
N1 Precision Std	0.029095	0.027466	0.030022	0.03107
N1 Precision 95% CI	[0.2648, 0.3781]	[0.2057, 0.3165]	[0.2475, 0.3686]	[0.4347, 0.5602]
N1 Recall Mean	0.715238	0.651407	0.742201	0.525901
N1 Recall Std	0.040557	0.043888	0.040263	0.052976
N1 Recall 95% CI	[0.6318, 0.7972]	[0.5615, 0.7308]	[0.6561, 0.8107]	[0.4204, 0.6289]
N1 F1 Mean	0.441066	0.371563	0.435678	0.508348
N1 F1 Std	0.028216	0.030217	0.031217	0.031979
N1 F1 95% CI	[0.3853, 0.4952]	[0.3056, 0.4268]	[0.3700, 0.4926]	[0.4418, 0.5676]
N2 Precision Mean	0.913831	0.858727	0.893689	0.835921
N2 Precision Std	0.012944	0.019862	0.018874	0.021796
N2 Precision 95% CI	[0.8879, 0.9388]	[0.8196, 0.8959]	[0.8550, 0.9280]	[0.7934, 0.8777]
N2 Recall Mean	0.737395	0.744479	0.71251	0.861961
N2 Recall Std	0.023285	0.032416	0.03283	0.036228

N2 Recall 95% CI	[0.6891, 0.7776]	[0.6768, 0.8025]	[0.6400, 0.7687]	[0.7833, 0.9187]
N2 F1 Mean	0.815925	0.79697	0.792396	0.848166
N2 F1 Std	0.014733	0.019147	0.022244	0.020071
N2 F1 95% CI	[0.7843, 0.8409]	[0.7575, 0.8303]	[0.7412, 0.8312]	[0.8027, 0.8816]
N3 Precision Mean	0.633804	0.5826	0.602463	0.847871
N3 Precision Std	0.057272	0.071601	0.057769	0.041691
N3 Precision 95% CI	[0.5163, 0.7393]	[0.4391, 0.7175]	[0.4829, 0.7093]	[0.7554, 0.9202]
N3 Recall Mean	0.911195	0.876297	0.931934	0.676726
N3 Recall Std	0.020433	0.024296	0.014907	0.044596
N3 Recall 95% CI	[0.8683, 0.9474]	[0.8280, 0.9260]	[0.9005, 0.9576]	[0.5858, 0.7601]
N3 F1 Mean	0.745891	0.696998	0.730108	0.751209
N3 F1 Std	0.040008	0.051947	0.043079	0.029535
N3 F1 95% CI	[0.6572, 0.8139]	[0.5843, 0.7855]	[0.6362, 0.8032]	[0.6877, 0.8051]
REM Precision Mean	0.769262	0.743989	0.81105	0.821325
REM Precision Std	0.03928	0.030373	0.024616	0.042184
REM Precision 95% CI	[0.6873, 0.8364]	[0.6849, 0.8028]	[0.7539, 0.8517]	[0.7317, 0.8918]
REM Recall Mean	0.823735	0.656873	0.769481	0.785393
REM Recall Std	0.023197	0.041032	0.027151	0.02964
REM Recall 95% CI	[0.7778, 0.8659]	[0.5759, 0.7351]	[0.7120, 0.8222]	[0.7256, 0.8398]
REM F1 Mean	0.794812	0.696601	0.789327	0.802318
REM F1 Std	0.022588	0.024819	0.019328	0.028335
REM F1 95% CI	[0.7448, 0.8349]	[0.6452, 0.7422]	[0.7494, 0.8235]	[0.7422, 0.8493]

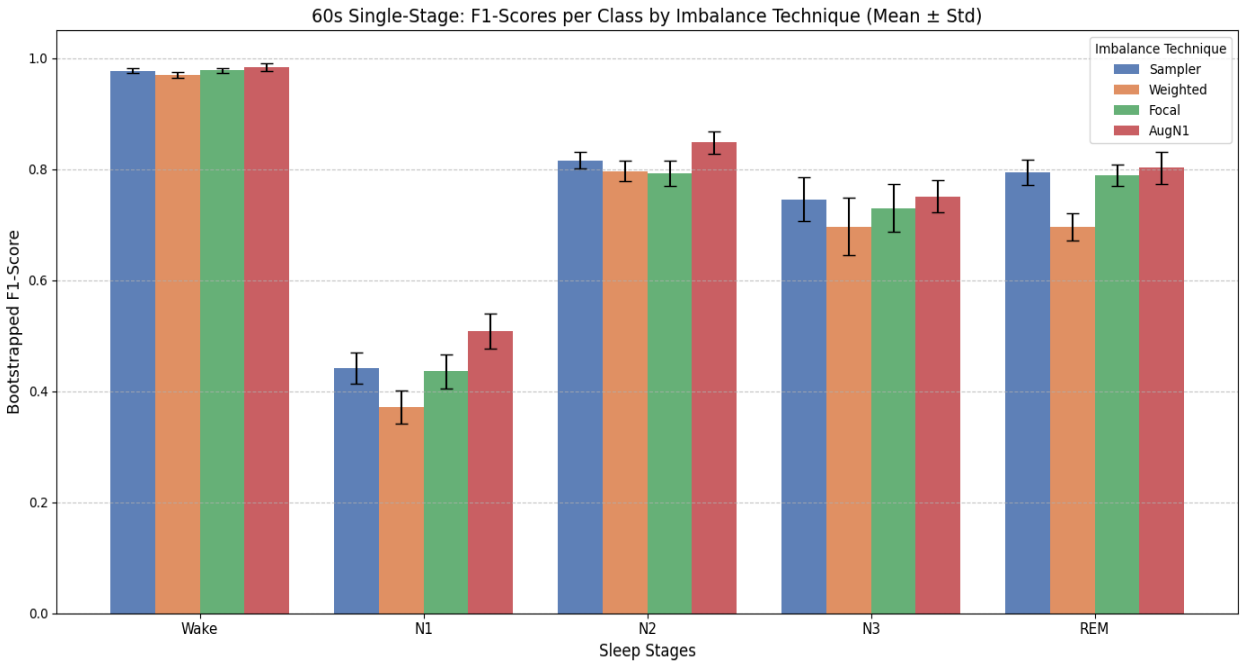

Figure 5. 60s 3ch Single Imbalance Technique Models

Table 6. 60s 3ch 2-stage Imbalance Technique Models

Model	**60s 2-Stage Sampler**	**60s 2-Stage Weighted**	**60s 2-Stage Focal**	**60s 2-Stage AugN1**
Accuracy Mean	0.915676	0.914873	0.910885	0.92045
Accuracy Std	0.014821	0.0145	0.014615	0.013742
Accuracy 95% CI	[0.8837, 0.9404]	[0.8843, 0.9399]	[0.8803, 0.9365]	[0.8917, 0.9442]
Macro Precision Mean	0.755623	0.748863	0.742414	0.774161
Macro Precision Std	0.014331	0.014328	0.012848	0.013698
Macro Precision 95% CI	[0.7245, 0.7815]	[0.7190, 0.7747]	[0.7152, 0.7654]	[0.7441, 0.7979]
Macro Recall Mean	0.788669	0.777733	0.782929	0.761814
Macro Recall Std	0.026822	0.025071	0.02561	0.025107
Macro Recall 95% CI	[0.7306, 0.8270]	[0.7242, 0.8170]	[0.7273, 0.8221]	[0.7067, 0.8023]
Macro F1 Mean	0.765002	0.758265	0.75173	0.76594

Macro F1 Std	0.017992	0.017841	0.016501	0.017488
Macro F1 95% CI	[0.7254, 0.7952]	[0.7171, 0.7883]	[0.7146, 0.7796]	[0.7266, 0.7953]
Wake Precision Mean	0.976564	0.976564	0.976564	0.976564
Wake Precision Std	0.015466	0.015466	0.015466	0.015466
Wake Precision 95% CI	[0.9417, 0.9952]	[0.9417, 0.9952]	[0.9417, 0.9952]	[0.9417, 0.9952]
Wake Recall Mean	0.985918	0.985918	0.985918	0.985918
Wake Recall Std	0.002885	0.002885	0.002885	0.002885
Wake Recall 95% CI	[0.9797, 0.9911]	[0.9797, 0.9911]	[0.9797, 0.9911]	[0.9797, 0.9911]
Wake F1 Mean	0.981156	0.981156	0.981156	0.981156
Wake F1 Std	0.008087	0.008087	0.008087	0.008087
Wake F1 95% CI	[0.9627, 0.9918]	[0.9627, 0.9918]	[0.9627, 0.9918]	[0.9627, 0.9918]
N1 Precision Mean	0.369984	0.413237	0.346274	0.430401
N1 Precision Std	0.032277	0.038795	0.032718	0.036608
N1 Precision 95% CI	[0.3106, 0.4394]	[0.3367, 0.4970]	[0.2833, 0.4143]	[0.3588, 0.5024]
N1 Recall Mean	0.564399	0.49265	0.559106	0.485572
N1 Recall Std	0.067285	0.062768	0.068042	0.062057
N1 Recall 95% CI	[0.4216, 0.6724]	[0.3622, 0.5996]	[0.4145, 0.6693]	[0.3584, 0.5984]
N1 F1 Mean	0.444529	0.446529	0.425223	0.453559
N1 F1 Std	0.03169	0.034655	0.032285	0.034603
N1 F1 95% CI	[0.3811, 0.5028]	[0.3734, 0.5086]	[0.3593, 0.4845]	[0.3810, 0.5146]
N2 Precision Mean	0.898977	0.88389	0.908528	0.845387
N2 Precision Std	0.015681	0.018692	0.01365	0.02006
N2 Precision 95% CI	[0.8680, 0.9292]	[0.8485, 0.9200]	[0.8824, 0.9334]	[0.8069, 0.8853]
N2 Recall Mean	0.762983	0.765148	0.74108	0.846094

N2 Recall Std	0.036036	0.034151	0.032309	0.031936
N2 Recall 95% CI	[0.6823, 0.8244]	[0.6921, 0.8249]	[0.6714, 0.7967]	[0.7774, 0.8976]
N2 F1 Mean	0.824833	0.819673	0.815872	0.845256
N2 F1 Std	0.021361	0.020072	0.020597	0.017634
N2 F1 95% CI	[0.7780, 0.8587]	[0.7765, 0.8539]	[0.7698, 0.8503]	[0.8057, 0.8746]
N3 Precision Mean	0.736441	0.681408	0.680193	0.795154
N3 Precision Std	0.049894	0.052503	0.049224	0.043409
N3 Precision 95% CI	[0.6322, 0.8238]	[0.5733, 0.7761]	[0.5805, 0.7676]	[0.6970, 0.8675]
N3 Recall Mean	0.856796	0.869119	0.893573	0.759384
N3 Recall Std	0.025367	0.034826	0.025169	0.04222
N3 Recall 95% CI	[0.8047, 0.9060]	[0.7938, 0.9334]	[0.8444, 0.9429]	[0.6790, 0.8393]
N3 F1 Mean	0.790772	0.762245	0.771005	0.77542
N3 F1 Std	0.028482	0.033526	0.030222	0.02722
N3 F1 95% CI	[0.7253, 0.8376]	[0.6888, 0.8198]	[0.7039, 0.8223]	[0.7151, 0.8227]
REM Precision Mean	0.79615	0.789217	0.800513	0.823299
REM Precision Std	0.043566	0.034252	0.03627	0.033519
REM Precision 95% CI	[0.7040, 0.8708]	[0.7116, 0.8447]	[0.7232, 0.8662]	[0.7468, 0.8773]
REM Recall Mean	0.77325	0.775829	0.734967	0.7321
REM Recall Std	0.03555	0.035704	0.03863	0.033754
REM Recall 95% CI	[0.7031, 0.8356]	[0.7020, 0.8397]	[0.6579, 0.8077]	[0.6602, 0.7916]
REM F1 Mean	0.783719	0.781721	0.765393	0.774308
REM F1 Std	0.030854	0.025708	0.026696	0.024643
REM F1 95% CI	[0.7160, 0.8350]	[0.7248, 0.8287]	[0.7071, 0.8133]	[0.7182, 0.8162]

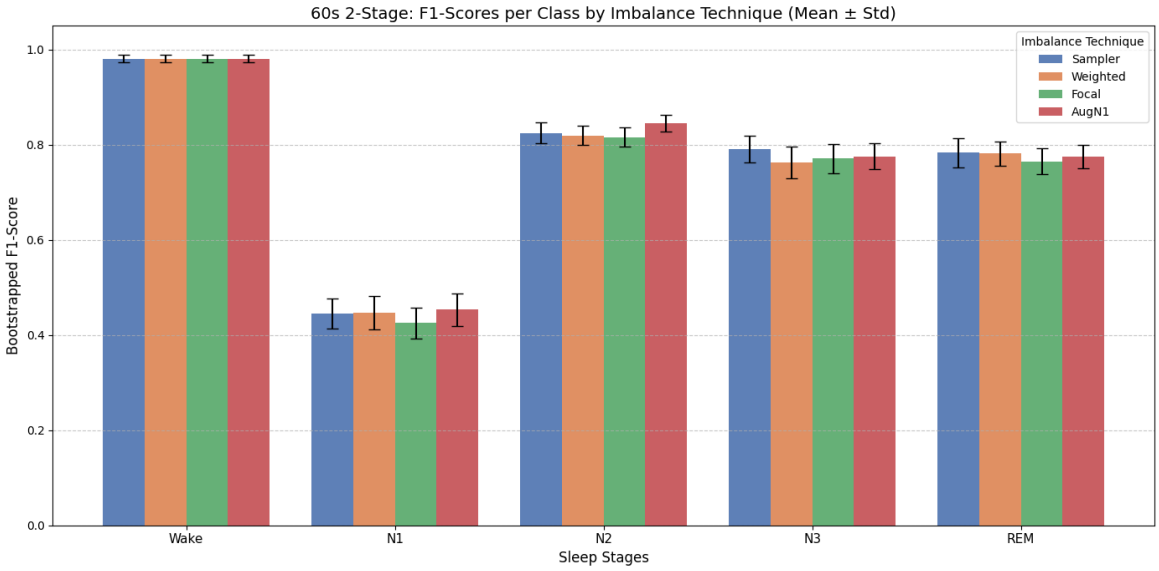

Figure 6. 60s 3ch 2-Stage Imbalance Technique Models

Table 7. 90s 3ch Single Imbalance Technique Models

Model	90s Single Sampler	90s Single Weighted	90s Single Focal	90s Single AugN1
Accuracy Mean	0.907239	0.873008	0.894512	0.930674
Accuracy Std	0.012047	0.013376	0.011798	0.00997
Accuracy 95% CI	[0.8838, 0.9301]	[0.8466, 0.8991]	[0.8722, 0.9173]	[0.9101, 0.9499]
Macro Precision Mean	0.737282	0.70247	0.703744	0.794549
Macro Precision Std	0.012964	0.01269	0.013422	0.01316
Macro Precision 95% CI	[0.7106, 0.7636]	[0.6760, 0.7267]	[0.6760, 0.7284]	[0.7668, 0.8182]
Macro Recall Mean	0.82895	0.791118	0.804695	0.782436
Macro Recall Std	0.013061	0.011181	0.014527	0.014638
Macro Recall 95% CI	[0.8013, 0.8533]	[0.7675, 0.8130]	[0.7754, 0.8318]	[0.7523, 0.8103]
Macro F1 Mean	0.761215	0.71215	0.733652	0.785926

Macro F1 Std	0.013133	0.011682	0.012906	0.012322
Macro F1 95% CI	[0.7345, 0.7867]	[0.6906, 0.7346]	[0.7073, 0.7575]	[0.7600, 0.8079]
Wake Precision Mean	0.997022	0.997123	0.996692	0.987603
Wake Precision Std	0.001202	0.000892	0.00128	0.003686
Wake Precision 95% CI	[0.9943, 0.9990]	[0.9952, 0.9986]	[0.9938, 0.9988]	[0.9796, 0.9939]
Wake Recall Mean	0.965485	0.93362	0.955808	0.988007
Wake Recall Std	0.006203	0.010265	0.006527	0.002566
Wake Recall 95% CI	[0.9524, 0.9760]	[0.9119, 0.9519]	[0.9422, 0.9676]	[0.9824, 0.9927]
Wake F1 Mean	0.98099	0.964298	0.975811	0.987801
Wake F1 Std	0.003381	0.005545	0.003582	0.002461
Wake F1 95% CI	[0.9738, 0.9868]	[0.9526, 0.9743]	[0.9684, 0.9822]	[0.9828, 0.9922]
N1 Precision Mean	0.315181	0.208638	0.289587	0.447955
N1 Precision Std	0.031848	0.025675	0.028634	0.033091
N1 Precision 95% CI	[0.2529, 0.3762]	[0.1613, 0.2602]	[0.2323, 0.3476]	[0.3878, 0.5178]
N1 Recall Mean	0.737634	0.725245	0.624218	0.529718
N1 Recall Std	0.032623	0.042325	0.047791	0.042005
N1 Recall 95% CI	[0.6756, 0.8050]	[0.6440, 0.8062]	[0.5351, 0.7201]	[0.4567, 0.6143]
N1 F1 Mean	0.440428	0.322887	0.394316	0.483873
N1 F1 Std	0.031115	0.030336	0.028448	0.024558
N1 F1 95% CI	[0.3781, 0.4993]	[0.2632, 0.3808]	[0.3358, 0.4504]	[0.4360, 0.5331]
N2 Precision Mean	0.899868	0.890418	0.891621	0.847351

N2 Precision Std	0.018278	0.016265	0.019631	0.019306
N2 Precision 95% CI	[0.8625, 0.9337]	[0.8567, 0.9218]	[0.8520, 0.9292]	[0.8104, 0.8852]
N2 Recall Mean	0.741866	0.726034	0.716588	0.877224
N2 Recall Std	0.029605	0.025337	0.02217	0.021274
N2 Recall 95% CI	[0.6791, 0.7951]	[0.6710, 0.7692]	[0.6693, 0.7562]	[0.8305, 0.9114]
N2 F1 Mean	0.81282	0.799492	0.79426	0.861766
N2 F1 Std	0.018604	0.015204	0.014859	0.013684
N2 F1 95% CI	[0.7719, 0.8461]	[0.7673, 0.8259]	[0.7625, 0.8215]	[0.8330, 0.8855]
N3 Precision Mean	0.675269	0.677734	0.636884	0.848883
N3 Precision Std	0.056446	0.051302	0.052419	0.0389
N3 Precision 95% CI	[0.5597, 0.7843]	[0.5742, 0.7711]	[0.5302, 0.7335]	[0.7624, 0.9159]
N3 Recall Mean	0.873145	0.896366	0.900479	0.742433
N3 Recall Std	0.030107	0.020283	0.032745	0.046686
N3 Recall 95% CI	[0.8156, 0.9328]	[0.8567, 0.9349]	[0.8326, 0.9580]	[0.6486, 0.8270]
N3 F1 Mean	0.759691	0.770481	0.744373	0.790704
N3 F1 Std	0.034362	0.032252	0.035591	0.029377
N3 F1 95% CI	[0.6864, 0.8188]	[0.7036, 0.8253]	[0.6689, 0.8040]	[0.7270, 0.8417]
REM Precision Mean	0.799067	0.738438	0.703938	0.840955
REM Precision Std	0.024915	0.038793	0.034481	0.046677
REM Precision 95% CI	[0.7461, 0.8462]	[0.6567, 0.8045]	[0.6341, 0.7673]	[0.7411, 0.9112]
REM Recall Mean	0.82662	0.674325	0.826381	0.7748
REM Recall Std	0.02729	0.039803	0.026782	0.02692

REM Recall 95% CI	[0.7717, 0.8745]	[0.5980, 0.7489]	[0.7711, 0.8742]	[0.7164, 0.8239]
REM F1 Mean	0.812147	0.70359	0.7595	0.805487
REM F1 Std	0.017569	0.025294	0.021431	0.023806
REM F1 95% CI	[0.7769, 0.8447]	[0.6527, 0.7497]	[0.7158, 0.7986]	[0.7536, 0.8481]

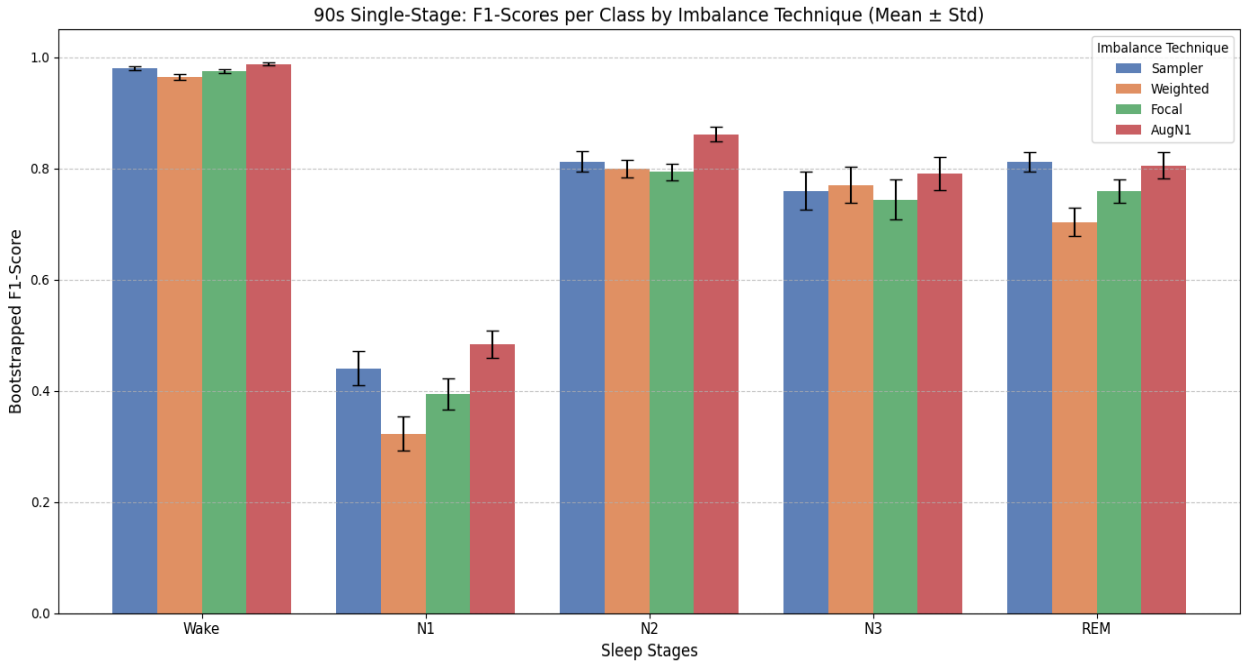

Figure 7. 90s 3ch Single Imbalance Technique Models

Table 8. 90s 3ch 2-stage Imbalance Technique Models

Model	90s 2-stage Sampler	90s 2-Stage Weighted	90s 2-Stage Focal	90s 2-Stage AugN1
Accuracy Mean	0.911566	0.909477	0.901021	0.918488
Accuracy Std	0.011736	0.012123	0.01227	0.010869
Accuracy 95% CI	[0.8889, 0.9348]	[0.8852, 0.9326]	[0.8773, 0.9254]	[0.8958, 0.9396]
Macro Precision Mean	0.735769	0.71702	0.69499	0.751592
Macro Precision Std	0.01306	0.013179	0.01151	0.014367

Macro Precision 95% CI	[0.7091, 0.7600]	[0.6897, 0.7417]	[0.6713, 0.7171]	[0.7223, 0.7790]
Macro Recall Mean	0.773873	0.746058	0.736048	0.733515
Macro Recall Std	0.014453	0.014906	0.013814	0.013882
Macro Recall 95% CI	[0.7441, 0.8002]	[0.7137, 0.7762]	[0.7086, 0.7633]	[0.7054, 0.7608]
Macro F1 Mean	0.743373	0.726974	0.708493	0.739595
Macro F1 Std	0.012389	0.013001	0.010919	0.012471
Macro F1 95% CI	[0.7170, 0.7667]	[0.6999, 0.7519]	[0.6865, 0.7298]	[0.7136, 0.7649]
Wake Precision Mean	0.986961	0.986961	0.986961	0.986961
Wake Precision Std	0.003838	0.003838	0.003838	0.003838
Wake Precision 95% CI	[0.9786, 0.9933]	[0.9786, 0.9933]	[0.9786, 0.9933]	[0.9786, 0.9933]
Wake Recall Mean	0.987303	0.987303	0.987303	0.987303
Wake Recall Std	0.002928	0.002928	0.002928	0.002928
Wake Recall 95% CI	[0.9811, 0.9923]	[0.9811, 0.9923]	[0.9811, 0.9923]	[0.9811, 0.9923]
Wake F1 Mean	0.987127	0.987127	0.987127	0.987127
Wake F1 Std	0.002564	0.002564	0.002564	0.002564
Wake F1 95% CI	[0.9817, 0.9917]	[0.9817, 0.9917]	[0.9817, 0.9917]	[0.9817, 0.9917]
N1 Precision Mean	0.297503	0.287159	0.268525	0.338611
N1 Precision Std	0.029836	0.034173	0.029292	0.036019
N1 Precision 95% CI	[0.2404, 0.3563]	[0.2205, 0.3593]	[0.2144, 0.3300]	[0.2749, 0.4154]
N1 Recall Mean	0.52902	0.412133	0.391238	0.400338
N1 Recall Std	0.058288	0.051115	0.062149	0.055522
N1 Recall 95% CI	[0.4116, 0.6361]	[0.3138, 0.5166]	[0.2687, 0.5095]	[0.2865, 0.5018]
N1 F1 Mean	0.378943	0.336447	0.316255	0.36421
N1 F1 Std	0.029426	0.031523	0.032533	0.032534

N1 F1 95% CI	[0.3215, 0.4325]	[0.2749, 0.4015]	[0.2527, 0.3765]	[0.2995, 0.4275]
N2 Precision Mean	0.881321	0.866035	0.877596	0.82792
N2 Precision Std	0.017766	0.020095	0.019571	0.020906
N2 Precision 95% CI	[0.8446, 0.9159]	[0.8244, 0.9044]	[0.8378, 0.9159]	[0.7846, 0.8683]
N2 Recall Mean	0.757127	0.769337	0.716565	0.86867
N2 Recall Std	0.022318	0.020767	0.018047	0.019981
N2 Recall 95% CI	[0.7110, 0.7983]	[0.7273, 0.8071]	[0.6830, 0.7522]	[0.8251, 0.9045]
N2 F1 Mean	0.814214	0.814482	0.788604	0.847494
N2 F1 Std	0.013777	0.01207	0.009274	0.012564
N2 F1 95% CI	[0.7863, 0.8412]	[0.7891, 0.8372]	[0.7700, 0.8073]	[0.8210, 0.8704]
N3 Precision Mean	0.700319	0.721902	0.697084	0.831284
N3 Precision Std	0.049611	0.052037	0.048604	0.044238
N3 Precision 95% CI	[0.5954, 0.7907]	[0.6079, 0.8144]	[0.5974, 0.7843]	[0.7369, 0.9103]
N3 Recall Mean	0.896727	0.840241	0.862897	0.719198
N3 Recall Std	0.021913	0.039773	0.029655	0.041762
N3 Recall 95% CI	[0.8487, 0.9370]	[0.7505, 0.9060]	[0.8066, 0.9211]	[0.6340, 0.8004]
N3 F1 Mean	0.785172	0.775101	0.769725	0.769717
N3 F1 Std	0.030524	0.034649	0.028486	0.027299
N3 F1 95% CI	[0.7198, 0.8390]	[0.7006, 0.8341]	[0.7092, 0.8161]	[0.7098, 0.8211]
REM Precision Mean	0.812742	0.723042	0.644784	0.773182
REM Precision Std	0.023065	0.028429	0.030568	0.04023
REM Precision 95% CI	[0.7657, 0.8557]	[0.6693, 0.7779]	[0.5882, 0.7048]	[0.6897, 0.8414]
REM Recall Mean	0.699187	0.721279	0.722238	0.692065
REM Recall Std	0.022562	0.026335	0.02909	0.033482

REM Recall 95% CI	[0.6529, 0.7427]	[0.6690, 0.7726]	[0.6625, 0.7771]	[0.6248, 0.7539]
REM F1 Mean	0.751408	0.721712	0.680757	0.729426
REM F1 Std	0.017587	0.020689	0.022985	0.025772
REM F1 95% CI	[0.7175, 0.7850]	[0.6808, 0.7602]	[0.6365, 0.7243]	[0.6807, 0.7774]

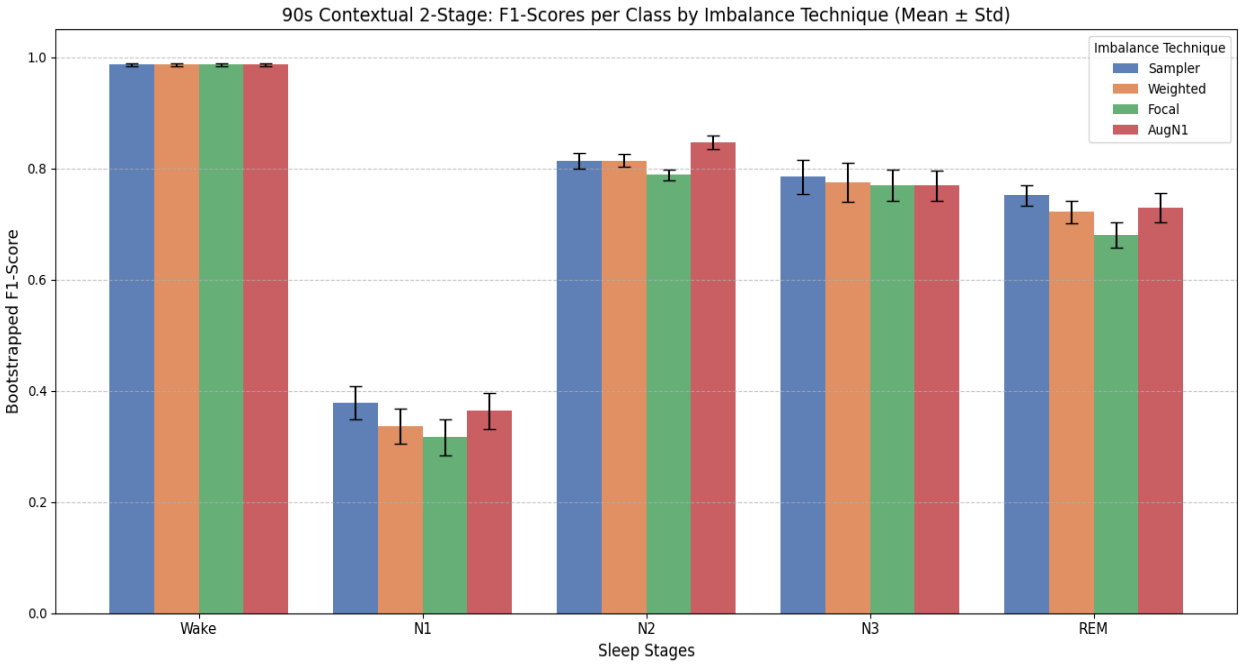

Figure 8. 90s 3ch 2-stage Imbalance Technique Models

Table 9. Wake Downsampled Buffer 30s Core Models

Model	**Buffer 2-Ch Single**	**Buffer 3-Ch Single**	**Buffer 3-Ch 2-Stage**
Accuracy Mean	0.911034	0.917874	0.91826
Accuracy Std	0.013213	0.011041	0.009375
Accuracy 95% CI	[0.8836, 0.9338]	[0.8965, 0.9381]	[0.8992, 0.9360]
Macro Precision Mean	0.723261	0.752714	0.75461
Macro Precision Std	0.017427	0.016372	0.013236
Macro Precision 95% CI	[0.6895, 0.7562]	[0.7182, 0.7826]	[0.7262, 0.7784]
Macro Recall Mean	0.696484	0.74001	0.747423
Macro Recall Std	0.014227	0.014662	0.0126

Macro Recall 95% CI	[0.6661, 0.7224]	[0.7112, 0.7698]	[0.7225, 0.7719]
Macro F1 Mean	0.703212	0.740536	0.745811
Macro F1 Std	0.013118	0.014874	0.012988
Macro F1 95% CI	[0.6763, 0.7277]	[0.7078, 0.7677]	[0.7172, 0.7697]
Wake Precision Mean	0.976711	0.98669	0.989486
Wake Precision Std	0.007215	0.003533	0.002364
Wake Precision 95% CI	[0.9604, 0.9880]	[0.9788, 0.9927]	[0.9846, 0.9937]
Wake Recall Mean	0.98698	0.979836	0.9751
Wake Recall Std	0.003328	0.004096	0.003913
Wake Recall 95% CI	[0.9797, 0.9925]	[0.9710, 0.9872]	[0.9672, 0.9822]
Wake F1 Mean	0.981804	0.983245	0.982235
Wake F1 Std	0.00413	0.002904	0.002443
Wake F1 95% CI	[0.9730, 0.9888]	[0.9775, 0.9886]	[0.9771, 0.9868]
N1 Precision Mean	0.363119	0.495434	0.497151
N1 Precision Std	0.038876	0.036247	0.036079
N1 Precision 95% CI	[0.2883, 0.4377]	[0.4211, 0.5665]	[0.4306, 0.5711]
N1 Recall Mean	0.186257	0.321133	0.331167
N1 Recall Std	0.017496	0.035353	0.046317
N1 Recall 95% CI	[0.1521, 0.2191]	[0.2580, 0.3921]	[0.2381, 0.4190]
N1 F1 Mean	0.245236	0.388294	0.39524
N1 F1 Std	0.018681	0.029786	0.036161
N1 F1 95% CI	[0.2099, 0.2806]	[0.3334, 0.4472]	[0.3182, 0.4635]
N2 Precision Mean	0.81945	0.837078	0.832807
N2 Precision Std	0.023769	0.021329	0.019684
N2 Precision 95% CI	[0.7749, 0.8653]	[0.7961, 0.8776]	[0.7963, 0.8700]
N2 Recall Mean	0.841798	0.855653	0.877168
N2 Recall Std	0.031233	0.02677	0.016426

N2 Recall 95% CI	[0.7768, 0.8944]	[0.7987, 0.9017]	[0.8432, 0.9055]
N2 F1 Mean	0.829896	0.845879	0.85419
N2 F1 Std	0.017059	0.016159	0.012079
N2 F1 95% CI	[0.7913, 0.8595]	[0.8099, 0.8747]	[0.8289, 0.8764]
N3 Precision Mean	0.757803	0.749898	0.776165
N3 Precision Std	0.051276	0.059976	0.049641
N3 Precision 95% CI	[0.6489, 0.8498]	[0.6187, 0.8559]	[0.6745, 0.8629]
N3 Recall Mean	0.751718	0.723293	0.77042
N3 Recall Std	0.048926	0.047825	0.038638
N3 Recall 95% CI	[0.6575, 0.8449]	[0.6332, 0.8205]	[0.6967, 0.8465]
N3 F1 Mean	0.752719	0.734115	0.771654
N3 F1 Std	0.032012	0.036074	0.027153
N3 F1 95% CI	[0.6805, 0.8100]	[0.6553, 0.7972]	[0.7150, 0.8189]
REM Precision Mean	0.699224	0.694469	0.67744
REM Precision Std	0.04399	0.043583	0.034996
REM Precision 95% CI	[0.6072, 0.7765]	[0.6077, 0.7731]	[0.6058, 0.7405]
REM Recall Mean	0.715667	0.820135	0.78326
REM Recall Std	0.037206	0.021731	0.025222
REM Recall 95% CI	[0.6388, 0.7884]	[0.7777, 0.8609]	[0.7324, 0.8317]
REM F1 Mean	0.706406	0.751149	0.725734
REM F1 Std	0.031777	0.026344	0.021197
REM F1 95% CI	[0.6377, 0.7597]	[0.6966, 0.7981]	[0.6818, 0.7641]

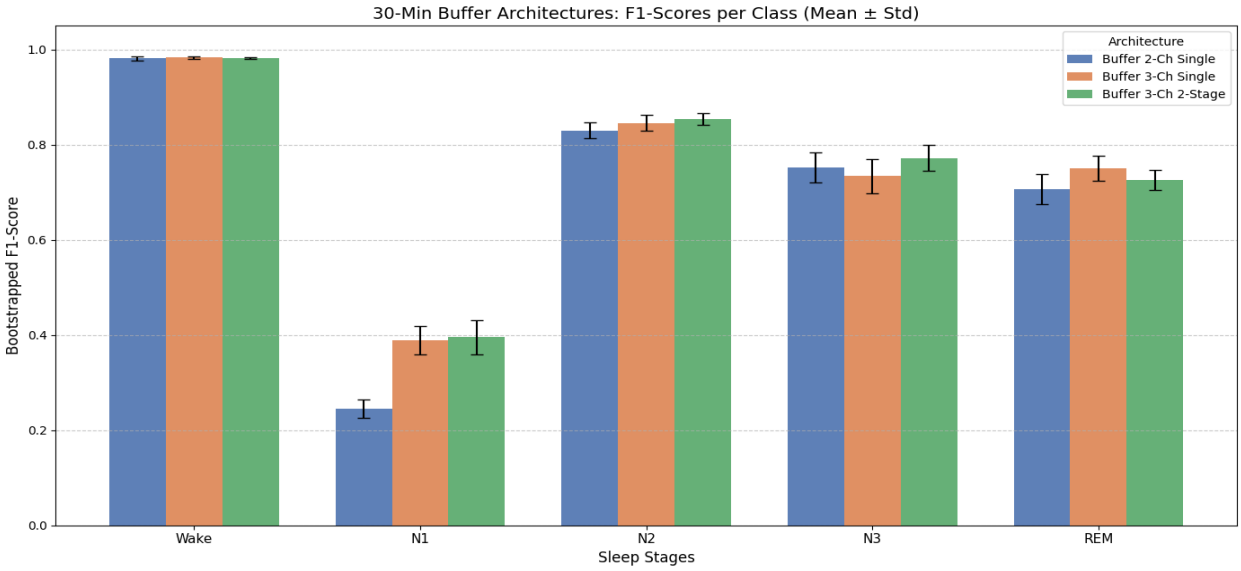

Figure 9. Wake Downsampled Buffer 30s Models

Discussion:

Under heavy class imbalance, the tested models were able to classify sleep stages at a clinical level, in which macro f1 ranged from 0.70-0.75 for the best models. It may be a consideration of whether the high ratio of wake to non-wake sleep stages in the dataset possibly weakened the model through encouraging a tendency to utilize a "baseline" route, i.e. the model tending towards defaulting to choosing "wake" due to its relative abundance. This was the reason a "buffer" model was tested, which utilized a "30-minute buffer" as conducted in preprocessing in contemporary literature, and was compared against this study's base 30s 3ch model. Both model's Cohen's kappa was around 0.83, suggesting that the high-class imbalance did not influence the models in this study to simply default to picking a majority, and instead learned the patterns and characteristics of each sleep stage.

The N1 stage was the weakest performing sleep stage (Figures 2 to 9). N1 is highly underrepresented in data and difficult to classify, with even expert human scorers disagreeing on the exact bounds that constitute N1 [29]. Difficulty stems in part due to N1 being close feature-wise to the wake stage and being a transition stage between wake and N2, making classification difficult. Due to this fact, much of the difference in results and the reason for the adjustment of model architecture and use of imbalance techniques is aimed towards improving N1 performance.

The three-channel model performance was statistically significantly greater for N1 than the corresponding two-channel model for all of the tested temporal intervals (Figure 1). The superiority of a multi-modal approach has been established in literature [30], and this study supports the better performance of multiple channels in comparison to fewer, in this case particularly the two EEG channels performing worse than the two EEG channels plus EOG. EEG measures brain electrical activity, and so is essential for polysomnography and sleep stage classification [31]**.** EOG measures eye activity, and so since some sleep stages have characteristic features relating to eye activity (namely REM sleep), EOG can aid in classification. Due to the superiority of three channels, the subsequent models all utilized three channels.

The best tested temporal interval on average was the 60s three channel two-stage model and was statistically significantly greater in N1 performance compared to the 60s three channel single model. However, the 60s three channel single model was not statistically significantly different, so overall the 60s three channel approach was optimal. A base 30s model with no added context already is quite sufficient, but the 60s was able to have higher average f1 score for other sleep stages, and a statistically significant improvement in N1 (Table 1,2; Figure 2, $p < 0.05$). The relatively good performance of models without added modification suggests a possibility for accurate sleep stage classification in applications of lower computational capability, such as in wearable sleep trackers [30]. These findings suggest the importance of context in sleep stage classification, especially for underrepresented stages such as N1. Phan et al. indeed noted the limitations of a 30s window for sleep staging [35]. It is clear that context can be beneficial, but context being longer than optimal can cause a "smoothing" effect, that can lead to even less minority class representation in training (termed "minority class wash out") when tested in Perlsev et al.) [36]. The ideal "60s" temporal interval is supported by Sun et al. 2019 [33].

Generally, the imbalance technique of two-stage models of the same temporal interval had similar results that were not statistically significantly different in any sleep stage. The 30s did not see statistically significant differences between different imbalance techniques. 60s imbalance techniques implemented in a "single" classifier architecture had N1 data augmentation having statistically significant performance in N1 and N2 f1-scores (Figure 4) but had no statistically significant difference for two-stage (Figure 5). Similarly, the single 90s with N1 data augmentation had a statistically significantly higher performance for N2 than the other three models, and for N1, the N1 data augmentation the highest average f1 score, and was statistically significantly better than the "weighted loss" and "focal loss" imbalance techniques. Yet, the 90s two stage model did not have statistically significant differences between techniques (Figure 8). The N1 and N2 were clearly directly impacted by the implementation of certain techniques to rectify the effects of heavy class imbalance, particularly for N1. For models where a statistically significant improvement was seen, this may have been due to WeightedRandomSampler and N1 data augmentation oversampling N1 and creating slight variants of N1 data, respectively. By lessening the data imbalance, the techniques were able to support a more stable learning trajectory for the model. Weighted and focal loss, on the other hand, seemed to overly punish the model during training, and hence had greater recall but at the cost of lower precision, and overall, no change in f1 (Tables 3, 5, 7). Single models with an imbalance technique implemented and two stage models with or without a technique implemented are not statistically significantly different. Therefore, it is suggested that the two-stage bypass architecture can match the implementation of imbalance techniques or make their use superfluous. Additionally, the similarity of WeighedRandomSampler and N1 data augmentation, but the significantly higher computational requirement of data augmentation, makes WeightedRandomSampler an optimal choice in numerous situations, but the effect being reduced with fewer channels (i.e. single versus 2-stage).

The two-stage bypass system was created to see if sleep stages being classified as wake may be reduced, stemming from wake being distinctly different from non-wake sleep stages in general. However, due to the relative similarity of N1 to wake, it might have been classified as wake during the first stage binary classifier. So the benefits of fewer incorrect N2, N3, and REM classification would be less than expected. Still, for the 30s and 90s models, the 2-stage model had a greater average f1-score (5-fold cross validation). But as previously stated, the performance was not statistically significantly better.

It may be true that the model could have had more optimal parameters, such as weights for models utilizing imbalance techniques. Here, reasonable weights were attempted to be chosen. For example, gamma was set to 2.0 for the "focal loss" technique, as originally proposed in the RetinaNet paper [34]. This paper utilized the novel technique in improving dense object identification, namely the large imbalance between background and foreground exposure. However, the technique has been utilized for improving transition identification, as in Perslev et al. 2019 and Phan et al. 2022 [35], [36]. This study did not iterably define the gamma parameter, so as other studies have, investigation of specific parameters definitions can be conducted.

It is a limitation of the study that it could be possible that due to computational restraints and long training times, the models could have been more finely tuned and may have produced different or better results. Furthermore, there may be some possibility for bias in the dataset, but the data was from an accredited, public database widely used by many similar sleep stage classification models. Additionally, the subject-level K-fold cross validation and bootstrapping utilized further bolstered the results.

The number of channels could also be a significant limitation. While the two-stage architecture in some temporal intervals was shown to be statistically similar to some imbalance techniques, showing a possibility for less computational requirements. However, in regard to the number of channels, three channels were shown to be superior to two. Sekkel et al. notes that while this is true, the practical application is limited, since more channels translates to additional measuring equipment on a patient and henceforth being more demanding [25]. This study looked to investigate the effects of various model optimization methods, and this facet was not considered greatly, instead investigating effects in regard to computational limitations. Hence, single channel EEG is a notable direction for future research. Additionally, the impact and significance of certain imbalance techniques may grow in magnitude with fewer channels, as this study shows.

Conclusion:

This study investigated ASSM through a base CNN-GRU deep learning model architecture, and certain model adjustments, namely single versus two stage bypass architecture, temporal interval of training data, and use of imbalance techniques (Weighted Random Sampler, weighted loss, focal loss, data augmentation). The best model utilized 2 EEG and EOG channels, using a 60s temporal interval architecture two-stage bypass system with N1 data augmentation, which had a macro f1 of around 0.78 ± 0.02. The results suggest that sleep stage scoring can be made more efficient and accurate through machine learning, and deep learning techniques improved upon traditional machine techniques. A two-stage bypass model architecture alone was able to match the use of Weighted Random Sampler and data augmentation in a single model at 60s and 90s temporal intervals, while the 30s had no statistically significant difference. Additionally, the importance of context with higher temporal intervals is shown.

Future research may investigate different metrics for feature extraction and temporal analysis, minimize resource use, or use different signals. It is evident that a CNN-GRU or a similar feature extraction in addition to a temporally based model has great promise in sleep stage classification, due to the nature of the physiological signals EEG and EOG, and the importance of context. The exact metrics used in the CNN

could be adjusted to change what features are extracted. A different physiological signal, such as electromyogram (EMG) or electrocardiogram (ECG), would necessitate such an adjustment. In any case, future research may find varying results depending on these changes. Different metrics for CNN feature extraction and GRU focus could be applied along with these additions.

Resource use is another prominent field for future study. Since automated sleep stage scoring can be applied not only to clinical environments but also mobile/hand-held sleep measuring devices, minimizing resource use while maximizing model performance. For instance, Boe et al. (2019) already has investigated low-resource sleep stage classification using wireless wearable sensors [34]. Many different combinations of preprocessing, model training, and additional parameters henceforth present themselves as interesting and promising routes for future directions.

References:

[1] T. L. Baker, "Sleep apnea disorders. Introduction to sleep and sleep disorders," *Med. Clin. North Am.*, vol. 69, no. 6, pp. 1123–1152, Nov. 1985, doi: 10.1016/s0025-7125(16)30979-8.

[2] T. L. Baker, "Introduction to Sleep and Sleep Disorders," *Med. Clin. North Am.*, vol. 69, no. 6, pp. 1123–1152, Nov. 1985, doi: 10.1016/S0025-7125(16)30979-8.

[3] A. D. Laposky, J. Bass, A. Kohsaka, and F. W. Turek, "Sleep and circadian rhythms: Key components in the regulation of energy metabolism," *FEBS Lett.*, vol. 582, no. 1, pp. 142–151, Jan. 2008, doi: 10.1016/j.febslet.2007.06.079.

[4] N. A. Jessen, A. S. F. Munk, I. Lundgaard, and M. Nedergaard, "The Glymphatic System: A Beginner's Guide," *Neurochem. Res.*, vol. 40, no. 12, pp. 2583–2599, Dec. 2015, doi: 10.1007/s11064-015-1581-6.

[5] M. M. Ohayon, "Epidemiological Overview of sleep Disorders in the General Population," *Sleep Med. Res.*, vol. 2, no. 1, pp. 1–9, Apr. 2011, doi: 10.17241/smr.2011.2.1.1.

[6] N. Pennings, L. Golden, K. Yashi, J. Tondt, and H. E. Bays, "Sleep-disordered breathing, sleep apnea, and other obesity-related sleep disorders: An Obesity Medicine Association (OMA) Clinical Practice Statement (CPS) 2022," *Obes. Pillars*, vol. 4, p. 100043, Dec. 2022, doi: 10.1016/j.obpill.2022.100043.

[7] 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, Jan. 2023, doi: 10.3390/app132413280.

[8] A. Rechtschaffen and A. Kales, "A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects".

[9] M. H. Silber *et al.*, "The Visual Scoring of Sleep in Adults," *J. Clin. Sleep Med.*, vol. 3, no. 2, pp. 121–131, Mar. 2007, doi: 10.5664/jcsm.26814.

[10] K. E. Bloch, "Polysomnography: a systematic review*," *Technol. Health Care*, vol. 5, no. 4, pp. 285–305, Jul. 1997, doi: 10.3233/THC-1997-5403.

[11] I. G. Campbell, "EEG Recording and Analysis for Sleep Research," *Curr. Protoc. Neurosci.*, vol. 49, no. 1, p. 10.2.1-10.2.19, 2009, doi: 10.1002/0471142301.ns1002s49.

[12] A. Morley and A. Morley, "10-20 system EEG Placement".

[13] B. Jammes, H. Sharabty, and D. Esteve, "Automatic EOG analysis: A first step toward automatic drowsiness scoring during wake-sleep transitions," *Somnologie - Schlafforschung Schlafmed.*, vol. 12, no. 3, pp. 227–232, Sep. 2008, doi: 10.1007/s11818-008-0351-y.

[14] "The electrooculogram," in *Handbook of Clinical Neurology*, vol. 160, Elsevier, 2019, pp. 495–499. doi: 10.1016/B978-0-444-64032-1.00033-3.

[15] M. Shokrollahi and S. Krishnan, "Sleep EMG analysis using sparse signal representation and classification," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2012, pp. 3480–3483. doi: 10.1109/EMBC.2012.6346715.

[16] K. Kesper, S. Canisius, T. Penzel, T. Ploch, and W. Cassel, "ECG signal analysis for the assessment of sleep-disordered breathing and sleep pattern," *Med. Biol. Eng. Comput.*, vol. 50, no. 2, pp. 135–144, Feb. 2012, doi: 10.1007/s11517-011-0853-9.
[17] P. Liu *et al.*, "Automatic sleep stage classification using deep learning: signals, data representation, and neural networks," *Artif. Intell. Rev.*, vol. 57, no. 11, p. 301, Sep. 2024, doi: 10.1007/s10462-024-10926-9.
[18] "AASM Scoring Manual - American Academy of Sleep Medicine," American Academy of Sleep Medicine – Association for Sleep Clinicians and Researchers. Accessed: Feb. 16, 2026. [Online]. Available: https://aasm.org/clinical-resources/scoring-manual/
[19] Y. Wu, Y. Lo, and Y. Yang, "STCN: A Lightweight Sleep Staging Model with Multiple Channels," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2020, pp. 1180–1183. doi: 10.1109/BIBM49941.2020.9313371.
[20] K. A. I. Aboalayon, M. Faezipour, W. S. Almuhammadi, and S. Moslehpour, "Sleep Stage Classification Using EEG Signal Analysis: A Comprehensive Survey and New Investigation," *Entropy*, vol. 18, no. 9, p. 272, Sep. 2016, doi: 10.3390/e18090272.
[21] E. Alickovic and A. Subasi, "Ensemble SVM Method for Automatic Sleep Stage Classification," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 6, pp. 1258–1265, Jun. 2018, doi: 10.1109/TIM.2018.2799059.
[22] P.-Y. Tsai, W. Hu, T. B. J. Kuo, and L.-Y. Shyu, "A portable device for real time drowsiness detection using novel active dry electrode system," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Sep. 2009, pp. 3775–3778. doi: 10.1109/IEMBS.2009.5334491.
[23] S. Yu, X. Chen, B. Wang, and X. Wang, "Automatic sleep stage classification based on ECG and EEG features for day time short nap evaluation," in *Proceedings of the 10th World Congress on Intelligent Control and Automation*, Jul. 2012, pp. 4974–4977. doi: 10.1109/WCICA.2012.6359421.
[24] Y. Li, F. Yingle, L. Gu, and T. Qinye, "Sleep stage classification based on EEG Hilbert-Huang transform," in *2009 4th IEEE Conference on Industrial Electronics and Applications*, May 2009, pp. 3676–3681. doi: 10.1109/ICIEA.2009.5138842.
[25] R. N. Sekkal, F. Bereksi-Reguig, D. Ruiz-Fernandez, N. Dib, and S. Sekkal, "Automatic sleep stage classification: From classical machine learning methods to deep learning," *Biomed. Signal Process. Control*, vol. 77, p. 103751, Aug. 2022, doi: 10.1016/j.bspc.2022.103751.
[26] B. Kemp, Aeilko Zwinderman, B. Tuk, H. Kamphuisen, and J. Oberyé, "The Sleep-EDF Database [Expanded]." physionet.org, 2018. doi: 10.13026/C2X676.
[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.
[28] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. doi: 10.3115/v1/D14-1179.
[29] J. Zhang and Y. Wu, "A New Method for Automatic Sleep Stage Classification," *IEEE Trans. Biomed. Circuits Syst.*, vol. 11, no. 5, pp. 1097–1110, Oct. 2017, doi: 10.1109/TBCAS.2017.2719631.
[30] M. Younes, J. Raneri, and P. Hanly, "Staging Sleep in Polysomnograms: Analysis of Inter-Scorer Variability," *J. Clin. Sleep Med.*, vol. 12, no. 6, pp. 885–894, Jun. 2016, doi: 10.5664/jcsm.5894.
[31] X. Zhang, G. He, T. Shang, and F. Fan, "Comparative Analysis of Single-Channel and Multi-Channel Classification of Sleep Stages Across Four Different Data Sets," *Brain Sci.*, vol. 14, no. 12, p. 1201, Nov. 2024, doi: 10.3390/brainsci14121201.
[32] Z. Liang and M. A. Chapa-Martell, "A Multi-Level Classification Approach for Sleep Stage Prediction With Processed Data Derived From Consumer Wearable Activity Trackers," *Front. Digit. Health*, vol. 3, May 2021, doi: 10.3389/fdgth.2021.665946.
[33] C. Sun, C. Chen, W. Li, J. Fan, and W. Chen, "A Hierarchical Neural Network for Sleep Stage Classification Based on Comprehensive Feature Learning and Multi-Flow Sequence Learning," IEEE J. Biomed. Health Inform., vol. 24, no. 5, pp. 1351–1366, May 2020, doi: 10.1109/JBHI.2019.2937558.
[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," Feb. 07, 2018, *arXiv*: arXiv:1708.02002. doi: 10.48550/arXiv.1708.02002.

[35] H. Phan and K. Mikkelsen, “Automatic sleep staging of EEG signals: recent development, challenges, and future directions,” *Physiol. Meas.*, vol. 43, no. 4, p. 04TR01, Apr. 2022, doi: 10.1088/1361-6579/ac6049.
[36] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, “U-Sleep: resilient high-frequency sleep staging,” *Npj Digit. Med.*, vol. 4, no. 1, p. 72, Apr. 2021, doi: 10.1038/s41746-021-00440-5.
[37] A. J. Boe *et al.*, “Automating sleep stage classification using wireless, wearable sensors,” *Npj Digit. Med.*, vol. 2, no. 1, p. 131, Dec. 2019, doi: 10.1038/s41746-019-0210-1.