

Article

Non-Invasive Sleep Stage Classification with Imbalance-Aware Machine Learning for Healthcare Monitoring

Luisiana Sabbatini *, Alberto Belli , Sara Bruschi , Marco Esposito , Sara Raggiunto 
and Paola Pierleoni 

Department of Information Engineering (DII), Università Politecnica delle Marche, Via Brecce Bianche 12, 60131 Ancona, Italy; a.belli@univpm.it (A.B.); s.bruschi@staff.univpm.it (S.B.); m.esposito@staff.univpm.it (M.E.); s.raggiunto@staff.univpm.it (S.R.); p.pierleoni@univpm.it (P.P.)

* Correspondence: l.sabbatini@staff.univpm.it

Abstract

Sleep plays a fundamental role in human health and cognitive functioning, motivating the development of reliable and scalable methodologies for sleep stage classification (SSC). Recent advances in non-invasive and economically sustainable sensing technologies enable continuous sleep monitoring beyond laboratory settings. However, SSC remains a challenging data analytics task due to the intrinsic class imbalance among sleep stages. This study investigates the effectiveness of different imbalanced data management strategies within a machine learning framework for non-invasive SSC. The proposed approach relies exclusively on heart rate and motion signals, which can be acquired through wearable devices or contactless under-mattress sensors, making it suitable for longitudinal monitoring scenarios. Using the PhysioNet DREAMT dataset, 32 experimental scenarios are defined by combining data-level techniques (ADASYN oversampling with different balancing weights), algorithm-level strategies (cost-sensitive learning), and hybrid solutions. Four model families are evaluated—Decision Tree, k-Nearest Neighbors, Ensemble Classifiers, and Artificial Neural Networks—across classification tasks involving 2, 3, 4, and 5 sleep stages. The experimental results show that ensemble-based models provide robust and consistent performance under severe class imbalance, achieving macro accuracies of 82% for sleep–wake detection, 73% for 3-stage classification, 72% for 4-stage classification, and 64% for 5-stage classification. These findings confirm the relevance of imbalance-aware analytics and demonstrate the feasibility of accurate, minimally invasive SSC within big data and cognitive computing paradigms.

Keywords: sleep stage classification; imbalanced learning; non-invasive sensing; machine learning; data analytics



Academic Editor: Kostas Karpouzis

Received: 3 February 2026

Revised: 2 April 2026

Accepted: 7 April 2026

Published: 10 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Accurate monitoring and classification of sleep stages are critically important because sleep plays a fundamental role in regulating numerous physiological and cognitive processes, as highlighted by [1]. Sleep activity has been closely linked to immune function, metabolic regulation, memory consolidation, and emotional processing. Therefore, disrupted sleep patterns or misclassifications of sleep stages can hinder the timely identification of sleep-related disorders such as insomnia, sleep apnea, or narcolepsy. As such, the development of reliable computational models for sleep stage classification has significant implications not only for clinical diagnosis and personalized medicine, but also for broader efforts in preventive health and well-being monitoring.

In particular, adults who sleep less than seven hours a night are at greater risk of gaining weight, diabetes, high blood pressure, heart disease, stroke, and depression. However, the amount of sleep is certainly not the only important thing to monitor. According to [2], quality of sleep is indeed essential. Monitoring sleep stages during the night enables an objective assessment of a person's sleep quality. While sleeping, we spend different amounts of time in five different sleep stages (American Academy of Sleep Medicine—AASM-guidelines), as extensively described by [3]:

- Wake (W) refers to 5–15% of a healthy adult's night rest;
- Rapid eye movement (REM) refers to 20–25% of a healthy adult's night rest;
- Non-REM light sleep (NREM1) refers to 2–5% of a healthy adult's night rest;
- Non-REM medium sleep (NREM2) refers to 45–55% of a healthy adult's night rest;
- Non-REM deep sleep (NREM3) refers to 10–20% of a healthy adult's night rest.

Therefore, the physiology of sleep makes it an inherently imbalanced use case for data scientists aiming at the creation of SSC models.

The gold standard for monitoring sleep stages and associated sleep quality is polysomnography (PSG), which consists of electroencephalography (EEG), electromyography (EMG), electrocardiogram (ECG), and electrooculography (EOG) [4]. The raw PSG data are then manually annotated by professional technicians, who are able to estimate the sleep stage of monitored patients within 30 s epochs. The commonly accepted silver standard is actigraphy (ACG), but over time, its poor reliability and instability have been proven by [5]. The performance of the ACG was shown to be comparable or poorer than that of other wearable and sport tracking technologies, making the ACG not interesting for sleep stage classification (SSC). PSG is therefore the best and most reliable solution, but it is associated with high resource consumption, high costs, patient discomfort, and limited accuracy of interrater interpretation. Moreover, it is usually performed in specialized laboratories and is thus not viable for longitudinal monitoring [6].

Even though PSG is extremely reliable, it is expensive, invasive, complex, and not suitable for longitudinal studies. In this regard, the present work aims to answer the following research questions (RQ):

- (RQ1):** Is it possible to develop reliable and efficient modeling solutions able to directly perform SSC through non-invasive sensors able to measure HR and motion?
- (RQ2):** Do imbalance management techniques offer a way for improving achievable results in non-invasive SSC through HR and motion?

The remainder of this work is organized as follows: Section 2 deepens the literature focused on the two main topics of this work, non-invasive SSC and imbalance management techniques for imbalanced classification data. Section 3 is dedicated to a detailed presentation of the experiments performed and the dataset used. Furthermore, in Section 4, we present and discuss the achieved results. The last Section aims at summarizing the obtained results with respect to RQs and comparing our best model's performances to those of similar works in the current literature. Additionally, we summarize the findings and intuitions developed on the basis of the extensive number of experiments performed.

2. Related Works

Some researchers attempted to automate the annotation task, owing to the development of solutions capable of extracting features from raw ECG signals. They then used these predictors for the classification of sleep stages. In [7], HR, RR and motion were shown to be suitable for the SSC task. Two training scenarios (subject-specific and subject-independent) were explored. The results achieved are as follows: 51% accuracy in discerning all five sleep stages (5-SSC) and 77% accuracy in 3-SSC (NREM1-2-3 phases condensed into one

single NREM). Nonetheless, HR, BR and motion are extracted from raw PSG signals, thus being potentially different from the HR, BR and motion that other less intrusive monitoring technologies could provide. Similarly, ref. [8] used HR, RR, and movement (number of movements in a 30 s window) for SSC. These signals are extracted and computed from the original PSG signals. Ref. [9] used HR (directly derived from ECGs) data only for the SSC task. They perform 5-SSC with 66% accuracy, and 72% for 4-SSC (with NREM1 and NREM2 aggregated in a single state). Ref. [10] presented an open source Python 3.12 package for sleep staging based on heart rate variability that can be extracted from raw ECGs. Recent studies have also explored machine learning approaches for automated sleep analysis using physiological signals, highlighting the growing potential of data-driven techniques for unobtrusive sleep monitoring [11]. In [4], a generative adversarial network (GAN) was proposed for the management of class imbalance. The input of the network is the raw EEG signals from the acquired PSG. Nonetheless, the authors are not interested in non-invasive SSC.

2.1. Modern SSC Solutions

PSG weakness, together with advancements in sensing technologies and data analysis and modelling solutions, fostered the adoption of wearable devices [12] and noncontact sensors to build models capable of accurately estimating sleep stages, without the need for invasive and expensive PSG measurements [1]. Ref. [13] employed HR and motion count signals extracted from a wearable device. They compare several DL and ML models and some imbalance management techniques. They used an open dataset, which involved 31 subjects. They perform binary classification only (sleep-wake recognition), and the results are approximately 91–95% accuracy, 63–67% specificity, 94–98% sensitivity. Ref. [14] solved the 4-SSC task via a Fitbit device (wearable consumer tracker). Two sequential models are developed: the first cleans sleep stages computed by the proprietary Fitbit algorithm; in the case of misclassified sleep stage detection, the second model computes the correct sleep stage from raw data collected from the wearable device. The authors use random upsampling (RUS) and random downsampling (RDS). The effectiveness of under/oversampling has been investigated in another dedicated work [15]. Ref. [6] used an EarlySense contactless device. This work is focused on 4 SSC and sleep-wake (SW) recognition methods. Ref. [5] included in their study 7 consumer devices for sleep stage monitoring (4 wearables, 1 EarlySense, 2 based on the RF wave). A comparison with PSG is performed for both the Epoch-By-Epoch (EBE) 4 SSC and the aggregated sleep summary measures.

For the effective development of non-invasive SSC solutions based on ML and DL, the availability of open data is essential, and the recent review of [1] lists the datasets available up to 2022. Recently, an open dataset for sleep stage classification was published by [16] on the PhysioNet platform developed by [17]. It comprises 100 nights monitored via both the PSG and a wearable device (Empatica E4 bracelet), thus being a suitable reference for building and comparing data-driven modeling and data imbalance management solutions for sleep stage classification via simpler and less invasive monitoring technology. This open dataset (together with other open datasets) was used by [18], who focused on SW recognition through acceleration signals and self-supervised modeling.

In summary, much attention has been given to sleep monitoring over time; however, several studies have not focused on making it a non-invasive and easier task through the development of algorithms for wearable or noncontact device signals. Studies focused on assessing the reliability of contactless technologies (such as undermattress belts) are usually pilot studies with few patients involved (approximately 5 people). Among previous studies addressing our same objective, namely, developing algorithms based on easily collectable

physiological signals for accurate sleep stage classification, and using this framework to evaluate data imbalance management techniques, only a few have considered both multi-class sleep stage recognition beyond the simple sleep–wake (SW) task and the impact of imbalance mitigation strategies [13].

2.2. Class Imbalance Management Techniques

Sleep stage classification, akin to numerous other tasks in biomedical data analysis, is notably characterized by pronounced class imbalance. This intrinsic characteristic of data necessitates the rigorous assessment of mitigation strategies aimed at addressing data complexity and enhancing the robustness and generalizability of data-driven models in highly imbalanced and heterogeneous settings such as the one selected.

The class imbalance problem is evident in either simple sleep/wake recognition or up to more subtle sleep stage classification tasks focused on REM, NREM1 (also called light sleep), NREM2, NREM3 (also called deep sleep), and wake state recognition. Specifically, a normal human night's sleep typically consists of 40 to 60% of the time spent in light sleep (NREM 1 and 2), 15 to 25% of REM sleep, 15 to 20% of deep sleep (NREM 3), and 5 to 15% of wakefulness [19].

There are several domains of ML applications characterized by a strong class imbalance, especially if we focus on biomedical application scenarios. Several strategies have been proposed and adopted over time.

We can summarize the main typologies as follows:

- Data-level techniques—data resampling (over- and undersampling) such as the renowned synthetic minority oversampling technique (SMOTE);
- Algorithm-level techniques—cost-sensitive learning (CSL), and imbalance robust models, such as ensembles;
- Other approaches—hybrid combinations.

Ref. [20] proposed an overview of imbalanced data for the classification tasks in several domains. The authors distinguish between external solutions (which modify the data, but not the algorithms) and internal solutions (based on algorithms or training strategies that can manage imbalance). This review does not focus on multilabel classification and is mostly focused on binary classification, which is not the only task addressed in our study. A very complete article is the one by [21], where CSL, resampling, and algorithmic strategies are reviewed. The authors highlight the difficulty of setting the cost in CSL. Unfortunately, in their review discussion, they are only focused on binary classification. However, they delineate several research directions, and some have been explored in the meantime between their review and this paper. In [22], several resampling techniques (both undersampling and oversampling), and some imbalance-robust models (ensemble boosted, among others) are presented. Ref. [23] review is focused on multi-class medical tasks, which are characterized by a strong data imbalance. The authors focus on oversampling and provide a list of CSL and ensemble models capable of dealing with imbalanced data. Among the most recent contributions in this field, ref. [24] proposed a survey on oversampling (OS), undersampling (US), imbalance-robust algorithms, and CSL. This reference highlights the instability of the approaches tested and the need to carefully evaluate the effectiveness of imbalanced data management strategies case by case.

As briefly introduced, several strategies have been proposed to address class imbalance in machine learning. These methods can be broadly categorized into data-level approaches, algorithm-level approaches, and other complementary techniques.

2.2.1. Data-Level Imbalance-Handling Techniques

Ref. [25] explored undersampling, oversampling, and hybrid-sampling techniques. They compared 6 classifiers trained with 25 datasets, each characterized by different imbalance ratios, but mostly made up of few features and few samples. Moreover, they are focused only on binary classification. The effectiveness of SMOTE for a Parkinson's disease monitoring solution was tested in [26]. Ref. [27] described the adaptive synthetic (ADASYN) oversampling technique as an improvement over SMOTE. In fact, ADASYN creates instances in the interior of the minority class, using the weighted distribution for the minority classes instead of creating completely synthetic samples. The master's thesis by [28] focused only on binary classification. It deepens non-medical use cases by testing logistic regression (LR), support vector machines (SVM), random forest (RF) models with SMOTE or ADASYN or nothing. Their conclusions show that there is no preprocessing method that consistently improves the performance of trained models. In [29], a new oversampling technique is introduced: the Mahalanobis distance-based oversampling technique (MDO). The authors compare it with SMOTE and ADASYN via 20 multiclass datasets (3–26 classes each, 100–20 k samples each). MDO proved to be the best solution; however, this methodology has not received attention in the literature, which is why we decided to test the most commonly used ADASYN strategy.

2.2.2. Algorithm-Level Imbalance-Handling Techniques

Ref. [30] demonstrated the effectiveness of CSL: based on the experiments performed, the authors conclude that the reliability of CSL and resampling depends on the specific characteristics of the dataset and the application context. Ref. [31] demonstrated that four different ML models reach better performance with the CSL configuration using four different medical datasets. Ref. [32] reviewed CSL for medical data; however, resampling was not considered in their work, which highlights the importance of data and code sharing for this specific research topic. Furthermore, ref. [33] defined a new CSL method and validated its efficiency.

Analyzing hybrid techniques, the effectiveness of CSL coupled with oversampling to achieve better results has been demonstrated for bankruptcy modeling by [34]. Ref. [35] also used oversampling together with CSL on 4 non-biomedical datasets. The combination of these two approaches achieves better performance than both used separately. Ref. [36] tested single and ensemble classifiers, and ADASYN coupled with CSL. Nevertheless, they performed the experiments only on a proprietary dataset (regarding the freezing of gait in Parkinson's patients). Ref. [37] investigated the effects of feature selection in conjunction with oversampling and CSL. The experiments use six binary classification biomedical datasets characterized by extremely high dimensionality, which are not comparable to our modeling scenario.

2.2.3. Other Imbalance-Handling Techniques

A work proposing a novel approach for binary classification is [38], in which the proposed methodology is validated using 11 datasets. Ref. [39] demonstrated overfitting and noise resulting from the application of data-level approaches. Ref. [40] explored whether class imbalance (whose hindering effects on ML modeling are proven) also has a negative influence on DL models (multilayer perceptron and convolutional neural network). They used several datasets, some belonging to the biomedical domain, to test the hypothesis. They conclude that both MLP and CNN suffer from data imbalance, especially when the available dataset size is limited. In particular, MLP is highly affected, whereas the CNN is slightly less affected. Ref. [41] focused on AI-based approaches for imbalanced datasets. In terms of modern AI-based methods, GANs are unstable and not reliable if used with limited

data. However, resampling techniques such as SMOTE and its variations, or ADASYN, proved to be as efficient as being limited and inadequate, depending on the different use cases.

Over time, different solutions have been proposed to address imbalanced data in binary and multiclass applications, but a clear and unified strategy for successfully accomplishing this objective is lacking. Ref. [42] highlighted how a common and clear usage of resampling techniques in imbalanced datasets is missing. The datasets are sampled by the authors to create different datasets with different imbalance ratios and sizes. The normal 70/30% train/test split is used for each experiment. The authors conclude that SMOTE and ADASYN are similar if the amount of data is not large and if the imbalance ratio is high. Their focus is on binary classification only, but on the basis of the reviewed literature, we can deduce how much this holds for multiclass applications as well. Ref. [41] reviewed the literature on oversampling methodologies, summarizing the efficiency and limitations of different approaches. The authors highlight the need for tailored strategies in specific use cases, in line with the majority of academic works.

Based on the reviewed literature, we believe that the implementation of experiments based on ML modeling without considering more complex approaches based on data-consuming DL techniques, such as GANs, should be proven in the context of SSC. This will provide a clearer view of this specific use case in terms of minimally invasive SSC solutions obtained, considering imbalanced data management techniques.

The landscape emerging from this extended review on the management of imbalanced classification tasks is diverse. Several specific application domains have been explored over time, but most of the conclusions drawn by researchers highlight the dependency on specific characteristics of data, the instability of advanced DL-based solutions, and the consequent potential for additional research on this topic. Most of the reviewed literature agrees with the use-case specificity when evaluating the effectiveness and reliability of data-level resampling solutions and model-level cost-sensitive solutions.

3. Materials and Methods

As highlighted in the previous section, several strategies, focused on different conceptual levels of ML modeling, can be implemented to address imbalanced data classification. From doing nothing at all, given that imbalance is common in many practical use cases, to customizing the cost matrix used during model training, or oversampling minority classes, or selecting imbalance-robust supervised models.

In this work, we specify 32 different scenarios on the basis of the concatenation of class-imbalance management strategies in the face of different SSC tasks. Specifically, we are going to address sleep-wake (SW) recognition, 3-SSC (wake vs. REM vs. NREM), 4-SSC (wake vs. NREM1-2 vs. NREM3 vs. REM), and 5-SSC (all the sleep stages labeled by sleep technicians). The labeling of the five sleep stages is aligned with American Academy of Sleep Medicine (AASM) guidelines; nonetheless, the imbalance associated with the other tasks is peculiar and different from that associated with 5-SSL. Hence, we also present the results achieved for the other classification tasks.

We tested several strategies (each called a scenario), both simple and hybrid, in each of the aforementioned tasks. The extensive description of the experiments performed allows us to evaluate evidence of the effectiveness in dealing with the SSC use case.

We used Python 3.9 for data import and preparation, together with MATLAB 2024b, to facilitate the model training and comparison phases. Hereafter, we are going to present the data used for the experiments.

3.1. Dataset

The dataset we used for the experiments is the “Dataset for Real-time sleep stage Estimation using Multisensor Wearable Technology (DREAMT)” [16], which is available on the PhysioNet platform [17]. Data are accessible upon registration and signing of a data use agreement, thus making it an open access dataset. It is a novel (2024) and quite extensive repository, where the Empatica E4 wearable device (validated biomedical device, [43]) has been coupled with invasive PSG sensors to collect multiple signals, and the associated sleep stage is accurately estimated by sleep technicians on the basis of a PSG every 30 s window. Data were acquired from 100 subjects, both healthy and with any disease (mainly obstructive apnea and obesity). The data acquisition protocol is fixed for all participants.

For every participant’s night, Empatica E4 collects the following:

- The timestamp [s] (64 Hz);
- Blood volume pulse (BVP) derived from a photoplethysmography (PPG) sensor (64 Hz);
- Interbeat interval (IBI) [ms] derived from the PPG (64 Hz);
- Electrodermal activity (EDA) [μ S] of the galvanic skin response sensor (4 Hz);
- Skin temperature [$^{\circ}$ C] from the infrared thermopile sensor (4 Hz);
- Triaxial accelerometry (32 Hz);
- Heart rate (HR) [bpm] estimated from the BVP signal (1 Hz).

Owing to the synchronization of the acquisition systems, these data are aligned with those associated with the PSG to allow sleep stage annotation for each 30-s window.

Additional information about the dataset can be found in the original reference. The details reported here provide essential information to be noted in order to understand the process followed in this study. In this study, the dataset split was performed at the epoch level, meaning that samples from the same subject may appear in both the training and testing sets. Although subject-independent evaluation protocols (e.g., LOSO) are commonly adopted, their application in this dataset is challenging due to severe class imbalance and the absence of specific sleep stages in some subjects, which may lead to unstable evaluation conditions.

3.2. Preprocessing

We used the aforementioned dataset to create an analytical model that is able to estimate the correct sleep stage on the basis of HR and motion only, given that these physiological signals can be reliably acquired using contactless sleep monitoring technologies, such as under-mattress sensing systems. This choice was made to emulate the type of information typically available in unobtrusive long-term sleep monitoring scenarios.

The first preprocessing step is computing the magnitude of the three acceleration signals. Moreover, we filtered out the “preparation” and “missing” stages annotated, given their lack of consistency with our main aim, i.e., the development of a reliable SSC algorithm.

Feature Extraction: For all the remaining windows, we extracted some relevant and commonly adopted statistical time-domain features [44,45] from both the HR sequence and the motion signal computed. The extracted features for both HR and motion are those listed in Table 1:

Table 1. Extracted features and associated equations.

Feature	Equation
Mean	$\mu = \frac{1}{N} \sum_{i=1}^N x_i$
Standard Deviation	$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2}$
First Quartile (Q1)	value of x at 25 percentile
Median (Q2)	value of x at 50 percentile
Third Quartile (Q3)	value of x at 75 percentile
Minimum	$\min(x_1, x_2, \dots, x_N)$
Maximum	$\max(x_1, x_2, \dots, x_N)$
Root Mean Square (RMS)	$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2}$
Kurtosis	$Kurt(x) = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^4}{\sigma^4}$
Peak-to-Peak	$P2P = \max(x_i) - \min(x_i)$
Crest Factor	$CF = \frac{\max(x_i)}{RMS}$
Skewness	$Skew(x) = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^3}{\sigma^3}$

The resulting dataset available for successive experiments is composed of 80,091 samples (overall number of 30-s windows associated with relevant sleep stages) of the 24 extracted features (12 extracted from the HR signal, and 12 extracted from the motion signal).

The minority degree of minority classes depends on the different SSC tasks we address among SW, 3-SSC, 4-SSC, and 5-SSC. The number of samples for each sleep stage associated with every SSC task is summarized in Table 2.

Table 2. Number of windows for each sleep stage in each SSC task addressed.

SSC Task	Sleep Stage				
	Wake	REM	NREM 1	NREM 2	NREM 3
SW	20,113 (25%)		59,978 (75%)		
3-SSC	20,113 (25%)	8406 (11%)	51,572 (64%)		
4-SSC	20,113 (25%)	8406 (11%)	48,869 (61%)	2703 (3%)	
5-SSC	20,113 (25%)	8406 (11%)	8824 (11%)	40,045 (50%)	2703 (3%)

This dataset has been randomly split into 80/20% training/testing for the baseline scenarios, and into 60/40% training/testing for the data resampling-based scenarios. This different proportion is related to the fact that oversampling synthesizes minority class samples, thus increasing the amount of data available for model training, while the testing amount remains fixed. Hence, retaining more data for testing is fair for rebalancing-based scenarios.

Adaptive synthetic oversampling (ADASYN): Concerning oversampling, we reviewed the literature and identified several traditional and innovative methodologies. The effectiveness and reliability of most of them have been proven in specific scenarios only, which is why we decided to test one modern yet renowned approach, the adaptive synthetic (ADASYN) approach. It is based on generating new minority class samples on the basis of local density. With respect to the most cited and used SMOTE, ADASYN is better at emphasizing instances that are hardly classifiable. When oversampling, it is important to choose the best balancing ratio. The optimization of the ratio is highly specific, and no academic reference highlights a quantitative way to set the best rebalancing weight. For this reason, we defined different scenarios, each characterized by a specific percentage (either 25, 50 or 100%) of the gap between each minority class and the majority class to be

synthesized. Practically speaking, if the gap between the NREM1 class and the majority class (i.e., NREM2) is 20,000, in the ADA25 scenario 5000 NREM1 samples have been synthesized; in the ADA50 scenario, 10,000 NREM1 samples are generated; and in the ADA100 scenario the numerical gap between each minority class and the majority class has been filled (this is the case for absolutely balanced datasets resulting from the oversampling phase). Therefore, the selected ratios (25%, 50%, and 100%) represent increasing levels of re-balancing, allowing us to analyze the effect of mild, moderate, and full oversampling on model performance.

Customization of the misclassification cost matrix (CSL): Another strategy we explored is the CSL. The misclassification costs were defined by considering the class distribution observed in the dataset. In particular, higher penalties were assigned to errors involving minority classes in proportion to their relative frequency, so that the resulting cost matrices reflect the imbalance ratio among sleep stages rather than relying solely on expert-defined values. Specifically, we adopted the basic approach, which attributes the highest cost (an integer number) to the first minority class, the lower integer to the second minority class, and so on up to the majority class's cost, which is 1. The misclassification cost matrices used in all the _CSL scenarios for every task are presented in Tables 3–6.

Table 3. Cost matrix for the _CSL scenarios addressing the SW task (sleep and wake).

		Predicted Class	
		Sleep	Wake
True Class	sleep	0	1
	wake	2	0

Table 4. Cost matrix for the _CSL scenarios addressing the 3-SSC task (NREM, REM, and wake).

		Predicted Class		
		NREM	REM	Wake
True Class	NREM	0	1	1
	REM	3	0	3
	wake	2	2	0

Table 5. Cost matrix for the _CSL scenarios addressing the 4-SSC task (NREM12, NREM3, REM, and wake).

		Predicted Class			
		NREM12	NREM3	REM	Wake
True Class	NREM12	0	1	1	1
	NREM3	4	0	4	4
	REM	3	3	0	3
	wake	2	2	2	0

Table 6. Cost matrix for the _CSL scenarios addressing the 5-SSC task (NREM1, NREM2, NREM3, REM, and wake).

		Predicted Class				
		NREM1	NREM2	NREM3	REM	Wake
True Class	NREM1	0	3	3	3	3
	NREM2	1	0	1	1	1
	NREM3	5	5	0	5	5
	REM	3	3	3	0	3
	wake	2	2	2	2	0

3.3. Experimental Scenarios

The summary of experimental scenarios defined with respect to imbalance management techniques applied is presented hereafter. In the experiments, a compact notation is adopted to describe the different scenarios under evaluation. The term *base* refers to the use of the original, unbalanced dataset; the term *CSL* indicates that a cost matrix was applied during training in that scenario; the term *ADAx* indicates the use of ADASYN oversampling with *x* being the ratio of the gap filled between each minority class and the majority class in the training set (*x* being either 25, 50 or 100%). So, for example, in the scenario *ADA25_CSL*, ADASYN was applied to the training data, filling a quarter of the gap between each minority class and the majority class, applying then a cost matrix during training.

Specifically, for each SSC task addressed, we experiment with different strategies, as summarized below:

- **baseline**—use normalized features extracted from the original data (80% training and 20% testing);
- **base_CSL**—use normalized features extracted from the original data + customized misclassification cost matrix during model training (80% training and 20% testing);
- **ADA100**—use ADASYN oversampled normalized features (on the 60% used for training) up to a completely balanced dataset, and test the resulting models on 40% of the original data taken apart for testing;
- **ADA50**—use ADASYN oversampled normalized features (on the 60% used for training) filling half the gap between each minority class and the majority class, and test the resulting models on 40% of the original data taken apart for testing;
- **ADA25**—use ADASYN oversampled normalized features filling a quarter of the gap between each minority class and the majority class, and test the resulting models on 40% of the original data taken apart for testing;
- **ADA100_CSL**—as ADA100, but also setting a customized misclassification cost, and then testing the resulting models on 40% of the original data taken apart;
- **ADA50_CSL**—as ADA50, but also setting customized misclassification costs, and then testing the resulting models on 40% of the original data taken apart;
- **ADA25_CSL**—as ADA25, but also setting customized misclassification costs, and then testing the resulting models on 40% of the original data taken apart.

Within each of the experimental scenarios presented, four different models have been trained and compared with each other: decision tree (DT), k-nearest neighbour (KNN), ensemble (ENS), and deep artificial neural network (ANN). With respect to the models' hyperparameters, every experiment trained the optimized version by using 30 trials led by the Bayesian optimizer.

Owing to the proposed experimental setup, we are able to draw conclusions about several topics:

- compare the effectiveness and reliability of different imbalanced data management techniques (with/without CSL, with ADASYN at varying degrees of synthesis) for every SSC task and hence under different imbalance ratios;
- compare different models such as DT, KNN, and ANN, as long as the imbalance-robust ENS, under every scenario and every SSC task.

4. Results and Discussion

4.1. Performance Metrics

In line with the literature reviewed, we decided to include, together with well-known classification metrics computed for each class, such as accuracy, precision, sensitivity,

specificity, and F1 score [46], other metrics that proved to be robust against unbalanced classification. Specifically, we propose both class-specific metrics, such as specificity, sensitivity, precision, F1 score, geometric mean and Matthew's correlation coefficient (MCC) [47], and overall metrics, like imbalance accuracy metric (IAM) proposed by [48] and micro accuracy. Both the MCC and IAM range from -1 to 1 , and values close to 1 are desirable, as they mean better classification performance. Both metrics are more suitable with respect to the previously mentioned metrics in cases such as the one under analysis in this paper, i.e., imbalanced multiclass classification. MCC includes all the values of the confusion matrix in its formula, and it is sensitive to imbalance; IAM identifies asymmetric errors and can be adopted in unbalanced multiclass scenarios. Table 7 summarizes the detailed descriptions of the metrics, the associated formulas, and the category to which each belongs.

Table 7. Performance metrics used to measure the multiclass (C classes) experiments.

Metric	Formula/Definition	Type
True Positives	$TP_c =$ correctly c-classified instances of c-class	class
True Negatives	$TN_c =$ correctly not c-classified instances of other classes	class
False Positives	$FP_c =$ incorrectly c-classified instances of other classes	class
False Negatives	$FN_c =$ incorrectly not c-classified instances of c-class	class
Sensitivity	$Sens_c = \frac{TP_c}{TP_c + FN_c}$	class
Specificity	$Spec_c = \frac{TN_c}{TN_c + FP_c}$	class
Precision	$Prec_c = \frac{TP_c}{TP_c + FP_c}$	class
F1 Score	$F1_c = \frac{2 \cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c}$	class
Geometric Mean	$GM_c = \sqrt{Sensitivity_c \cdot Specificity_c}$	class
Matthew's Correl. Coef.	$MCC_c = \frac{TP_c \cdot TN_c - FP_c \cdot FN_c}{\sqrt{(TP_c + FP_c)(TP_c + FN_c)(TN_c + FP_c)(TN_c + FN_c)}}$	class
Micro Accuracy	$MicroAcc = \frac{\sum_c TP_c}{\sum_c (TP_c + FP_c)}$	global
Imbalance Accuracy Metric	$IAM = \frac{1}{C} \cdot \sum_{c=1}^C \frac{TP_c - \max(FP_c, FN_c)}{\max(TP_c + FP_c, TP_c + FN_c)}$	global

4.2. Experimental Setup

In this work, several experiments under diverse modeling scenarios, addressing different SSC tasks, were performed. Specifically, we performed experiments embracing three main variability dimensions: different models, different data-level techniques (scenarios), and different SSC tasks. The presentation of all the results could be overwhelming for the reader. For this reason, visually powerful bar charts and bump charts are presented hereafter. Extensive tables containing all the detailed numerical results can be found in the Appendix A, while the best results are going to be described in the text.

The bump charts highlight the ranking of different modeling scenario metrics, with respect to every class (horizontal axis) for every model trained (groups on the horizontal axis) in every SSC task addressed (one chart for each task). In this work, we consider performance metrics where the higher the value is, the better it is; we assign the 1st rank to the highest value (best performance), and the 8th rank to the lowest value.

On the other hand, bar charts highlight the magnitude of the value and the difference among the facts represented. It is easier to assess the effectiveness of different scenarios and models on the basis of the bump charts and bar charts presented below.

4.3. Comparative Analysis

To obtain a comprehensive view of the experiments performed, we present bump charts of $F1_c$ score (Figure 1), MCC_c (Figure 2), and GM_c (Figure 3) for every task addressed, with the aim of understanding whether one “best model” can be elected.

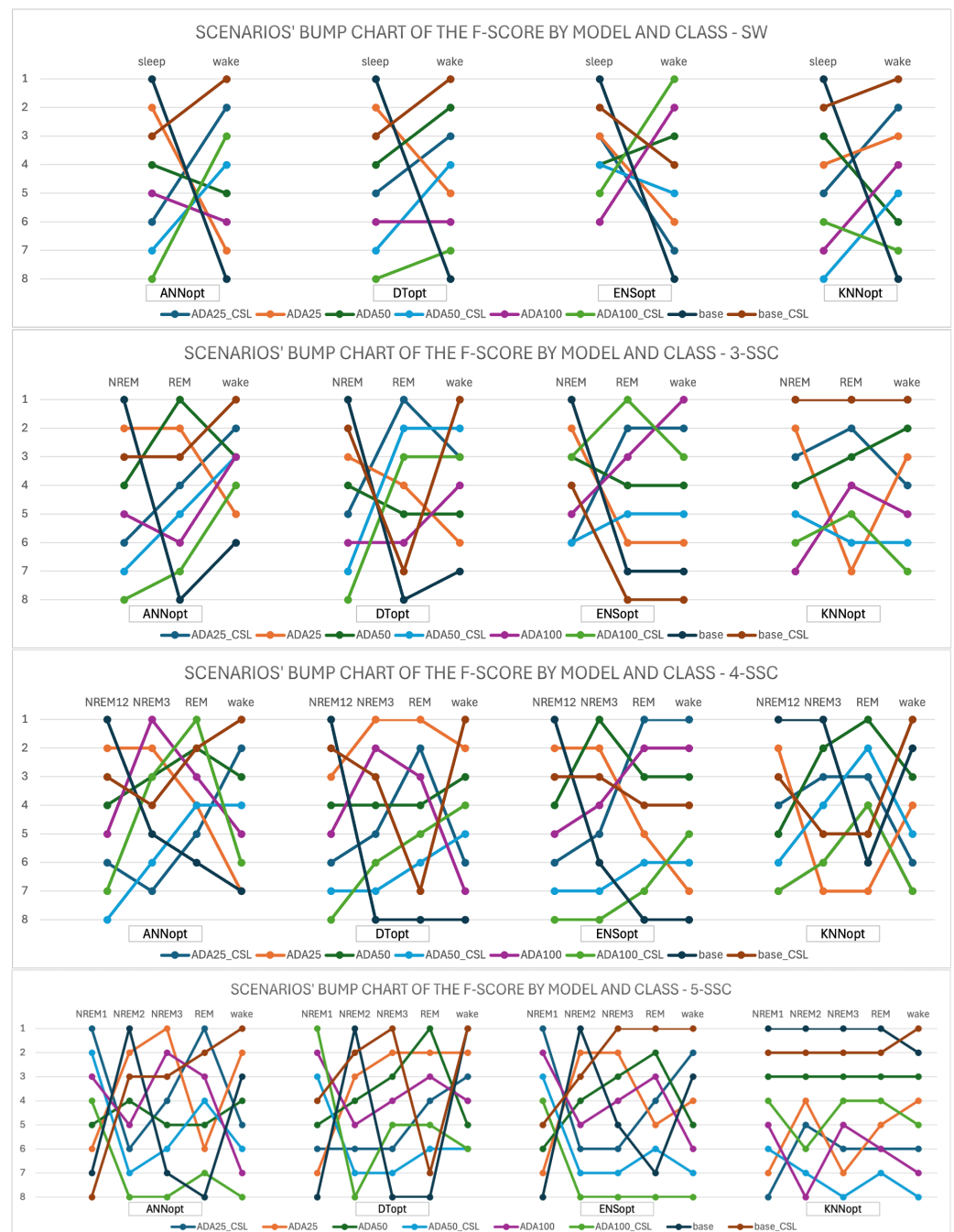


Figure 1. Bump charts (one for each task) ranking different scenarios' models $F1_c$ scores.

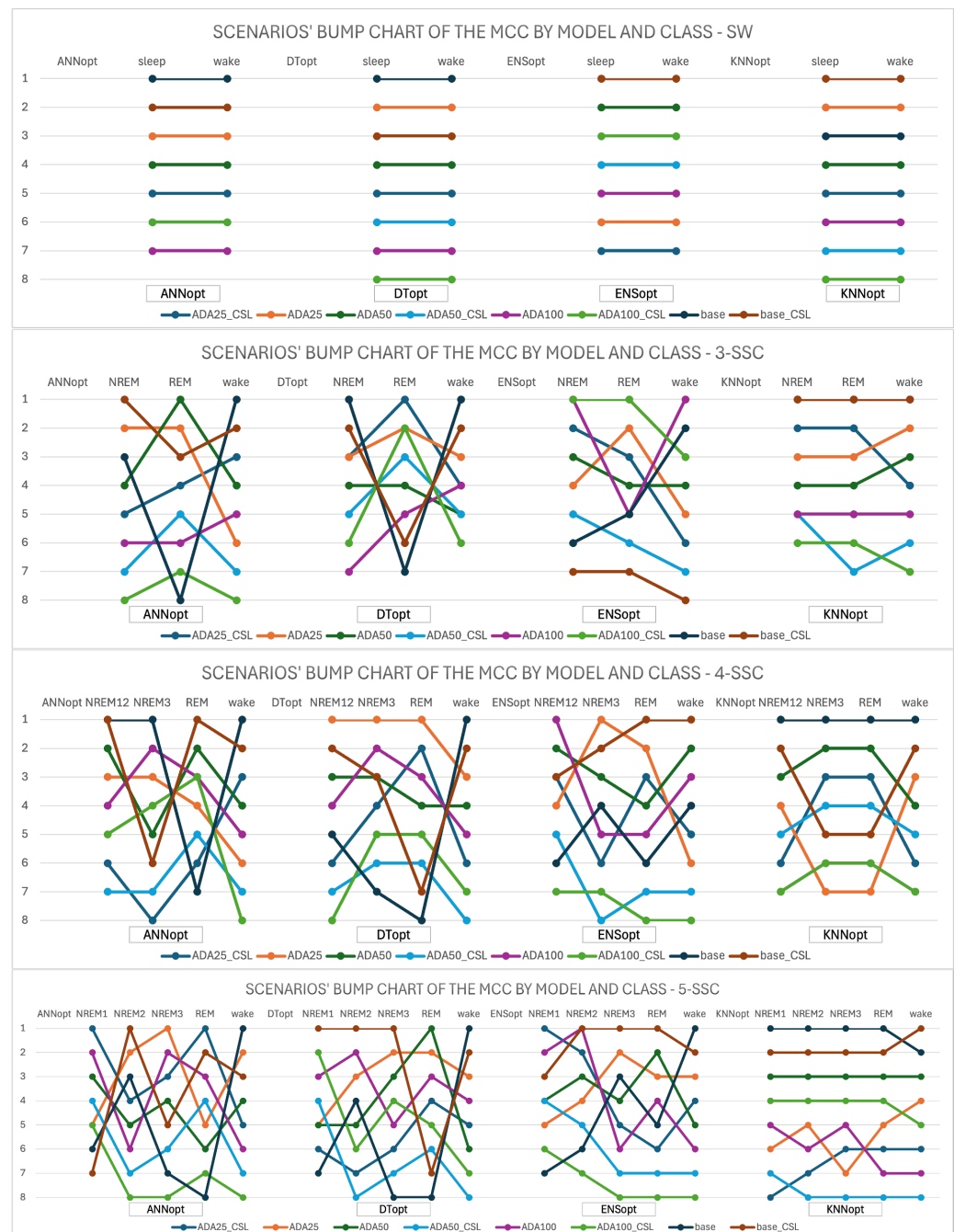


Figure 2. Bump charts (one for each task) ranking different scenarios' models MCC_c .

These bump charts allow us to deduce some general evidence:

- There is high variability in the results achieved under different SSC tasks; only the KNN model seems to have slightly more coherent behavior in the different classes (majority one and minority classes) for all three considered metrics.
- The $F1_c$ scores and MCC_c metrics are very coherent with each other as the number of classes increases.
- Different models appear to be differently affected by the imbalance management technique (oversampling and/or CSL); in simpler tasks (binary or ternary), ANN and DT are similar to each other, whereas in more complex tasks (4-SSC and 5-SSC) every model benefits from different modeling scenarios in a completely different way.

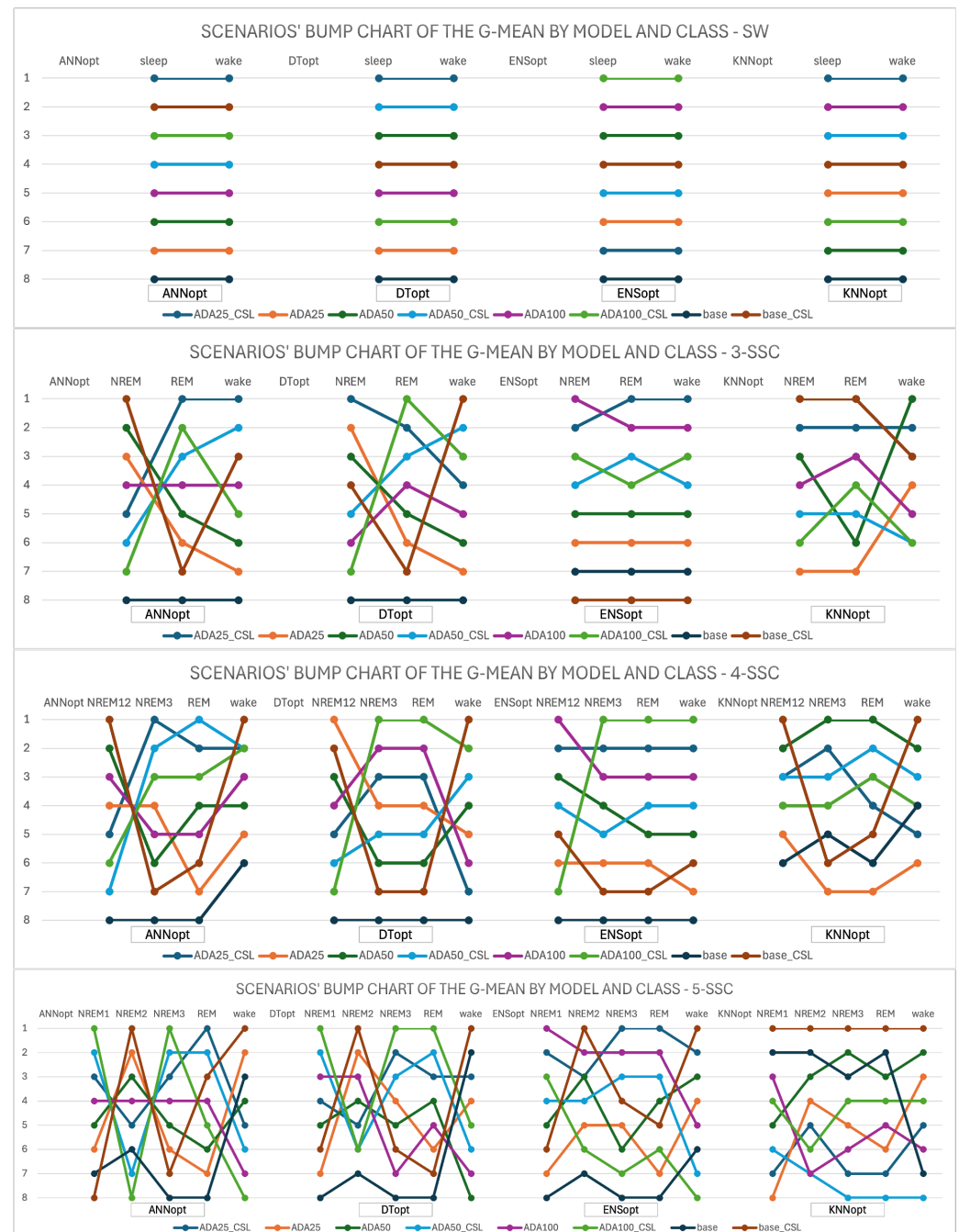


Figure 3. Bump charts (one for each task) ranking different scenarios' models GM_c .

Considering the previous results shown, the following plots provide a clearer focus on some specific points of interest. Figures 4 and 5 represent the IAM and micro accuracy values, respectively, for every model and scenario under the different tasks considered.

These allow us to deduce even deeper evidence in order to answer the RQs test:

- The more classes we try to detect, the worse the overall performance we obtain, as expected;
- Under every modeling scenario and classification task, the ensemble is usually the best-performing model;
- The superiority of the ENS over the DT, KNN classifier and ANN, is stable even in more complex multiclass tasks;
- By focusing on ENS, the base scenario together with the base_CSL and ADA_25 are usually the best scenarios.

- In general, every model usually performs best under the base, base_CSL and ADA_25 scenarios, suggesting that high weights for oversampling hamper the model performance of trained models.



Figure 4. Bar charts (one for each task) of the IAM metric value reached by each model (bar color) in every scenario (horizontal axis).

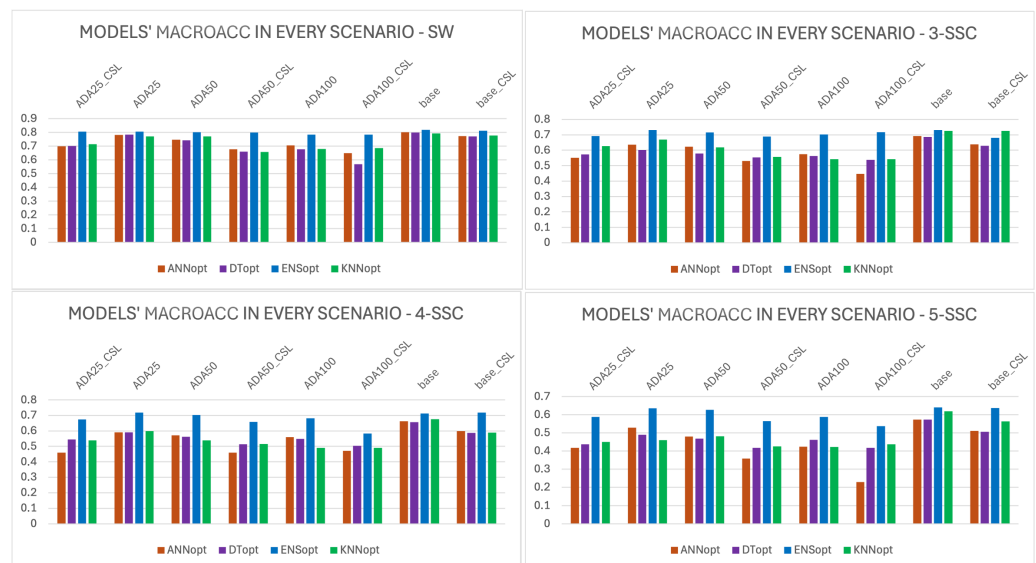


Figure 5. Bar charts (one for each task) of the Macro-Accuracy metric value reached by each model (bar color) in every scenario (horizontal axis).

4.4. Ensemble Model Analysis

To clarify the ensemble's superiority and dive deeper into the insights sketched before, we present heatmaps of the models' MCC_c metric (in green and bold the top values, in red lowest values) on the basis of scenario over model crossing for every class within each task addressed. Figures 6–9 show detailed results for each task addressed. The MCC_c metric proved to be the best metric in the case of imbalanced data, such as the use case considered in this study.

	ADA100	ADA100_CSL	ADA25	ADA25_CSL	ADA50	ADA50_CSL	base	base_CSL
sleep								
ANNopt	0.293	0.297	0.357	0.315	0.332	0.297	0.392	0.386
DTopt	0.257	0.231	0.366	0.311	0.338	0.285	0.385	0.364
ENSopt	0.43	0.436	0.429	0.428	0.448	0.431	0.455	0.455
KNNopt	0.312	0.267	0.371	0.337	0.347	0.285	0.354	0.385
wake								
ANNopt	0.293	0.297	0.357	0.315	0.332	0.297	0.392	0.386
DTopt	0.257	0.231	0.366	0.311	0.338	0.285	0.385	0.364
ENSopt	0.43	0.436	0.429	0.428	0.448	0.431	0.455	0.455
KNNopt	0.312	0.267	0.371	0.337	0.347	0.285	0.354	0.385

Figure 6. Heat-map of MCC_c values of considered models under different scenarios associated to the SW task (best value for each class in bold).

	ADA100	ADA100_CSL	ADA25	ADA25_CSL	ADA50	ADA50_CSL	base	base_CSL
NREM								
ANNopt	0.286	0.227	0.308	0.299	0.301	0.278	0.303	0.32
DTopt	0.242	0.252	0.267	0.267	0.257	0.254	0.283	0.28
ENSopt	0.448	0.448	0.424	0.437	0.427	0.419	0.403	0.257
KNNopt	0.219	0.212	0.292	0.317	0.289	0.219	0.444	0.444
REM								
ANNopt	0.224	0.201	0.249	0.235	0.256	0.225	0.128	0.24
DTopt	0.217	0.233	0.233	0.244	0.226	0.232	0.045	0.181
ENSopt	0.428	0.457	0.45	0.435	0.431	0.407	0.428	0
KNNopt	0.174	0.166	0.187	0.282	0.181	0.16	0.554	0.554
wake								
ANNopt	0.321	0.284	0.319	0.332	0.328	0.304	0.378	0.356
DTopt	0.263	0.256	0.27	0.263	0.261	0.261	0.362	0.331
ENSopt	0.444	0.441	0.437	0.436	0.439	0.422	0.442	0.346
KNNopt	0.24	0.227	0.344	0.308	0.326	0.236	0.387	0.387

Figure 7. Heat-map of MCC_c values of considered models under different scenarios associated with the 3-SSC task (best value for each class in bold).

	ADA100	ADA100_CSL	ADA25	ADA25_CSL	ADA50	ADA50_CSL	base	base_CSL
NREM12								
ANNopt	0.256	0.238	0.258	0.23	0.273	0.227	0.286	0.286
DTopt	0.26	0.233	0.285	0.247	0.261	0.237	0.253	0.275
ENSopt	0.441	0.391	0.428	0.433	0.44	0.415	0.401	0.433
KNNopt	0.174	0.174	0.214	0.202	0.238	0.204	0.326	0.258
NREM3								
ANNopt	0.398	0.389	0.394	0.308	0.381	0.335	0.401	0.353
DTopt	0.418	0.395	0.42	0.402	0.405	0.36	0	0.405
ENSopt	0.603	0.54	0.638	0.595	0.63	0.533	0.606	0.633
KNNopt	0.296	0.296	0.235	0.344	0.368	0.318	0.528	0.302
REM								
ANNopt	0.22	0.22	0.216	0.198	0.222	0.21	0.114	0.223
DTopt	0.253	0.238	0.299	0.263	0.244	0.22	0.046	0.206
ENSopt	0.441	0.379	0.488	0.455	0.45	0.419	0.421	0.498
KNNopt	0.143	0.143	0.094	0.162	0.186	0.161	0.237	0.157
wake								
ANNopt	0.305	0.275	0.298	0.329	0.319	0.295	0.371	0.363
DTopt	0.272	0.265	0.29	0.267	0.281	0.264	0.372	0.312
ENSopt	0.452	0.426	0.444	0.446	0.453	0.427	0.45	0.463
KNNopt	0.226	0.226	0.315	0.239	0.271	0.242	0.38	0.335

Figure 8. Heat-map of MCC_c values of considered models under different scenarios associated with the 4-SSC task (best value for each class in bold).

	ADA100	ADA100_CSL	ADA25	ADA25_CSL	ADA50	ADA50_CSL	base	base_CSL
NREM1								
ANNopt	0.093	0.069	0.069	0.103	0.092	0.09	0.016	0
DTopt	0.084	0.09	0.076	0.064	0.076	0.077	0.048	0.097
ENSopt	0.154	0.121	0.124	0.171	0.136	0.136	0.12	0.153
KNNopt	0.072	0.073	0.069	0.061	0.1	0.064	0.203	0.192
NREM2								
ANNopt	0.258	0.118	0.285	0.276	0.268	0.225	0.281	0.288
DTopt	0.3	0.277	0.289	0.271	0.28	0.261	0.284	0.305
ENSopt	0.451	0.381	0.435	0.449	0.442	0.42	0.419	0.451
KNNopt	0.237	0.256	0.248	0.229	0.308	0.21	0.439	0.433
NREM3								
ANNopt	0.37	0.242	0.394	0.363	0.346	0.31	0.289	0.335
DTopt	0.392	0.396	0.426	0.387	0.403	0.381	0.281	0.448
ENSopt	0.588	0.482	0.633	0.592	0.608	0.514	0.612	0.665
KNNopt	0.352	0.373	0.333	0.348	0.405	0.311	0.58	0.483
REM								
ANNopt	0.21	0.15	0.19	0.233	0.187	0.193	0.086	0.228
DTopt	0.257	0.242	0.269	0.247	0.274	0.239	0.065	0.202
ENSopt	0.447	0.35	0.464	0.443	0.468	0.42	0.446	0.488
KNNopt	0.178	0.2	0.187	0.182	0.244	0.153	0.588	0.531
wake								
ANNopt	0.307	0.295	0.367	0.308	0.341	0.299	0.37	0.342
DTopt	0.266	0.245	0.282	0.26	0.256	0.244	0.369	0.329
ENSopt	0.433	0.375	0.451	0.441	0.438	0.405	0.458	0.453
KNNopt	0.24	0.254	0.275	0.243	0.277	0.224	0.369	0.374

Figure 9. Heat-map of MCC_c values of considered models under different scenarios associated with the 5-SSC task (best value for each class in bold).

From the heatmaps, by looking at greener horizontal lines, it is evident again how much the ensemble model is the best performing model against imbalanced data classification under almost any imbalance management technique (either oversampling or/and CSL) for every class (either minority or majority).

Moreover, the vertical greener columns reveal how much the baseline and the baseline_CSL scenarios are generally the best ones in terms of the MCC, followed only by the ADA25 scenario, i.e., the scenario where a small number of minority classes' samples are synthesized. Nonetheless, the bold values (the best MCC for every class) range in cells linked to either ADA100, ADA100_CSL, ADA25, base, or base_CSL.

In summary, the ensemble is always the best model (coupled with variable scenarios) except for the following:

- REM (minority class in the 3-SSC task), where **KNN** is the best model;
- NREM1 (third minority class in the 5-SSC task), where **KNN** is the best model;
- REM (second minority class in the 5-SSC task), where **KNN** is the best model.

With respect to the scenarios, the landscape is more diverse:

- The **base** is the best (coupled with either Ensemble or KNN) in both classes for the SW task, in the minority class (REM) for the 3-SSC task, and for the REM, NREM1 and wake classes in the 5-SSC task;
- The **base_CSL** is the best (coupled with either Ensemble or KNN) in both classes for the SW task, in the minority class (REM) for the 3-SSC task, in the REM and wake classes for the 4-SSC task, and in the NREM2 and NREM3 classes for the 5-SSC task;
- The **ADA100** is the best (coupled with the ensemble only) in the NREM and wake classes for the 3-SSC task, in the NREM12 (majority) class for the 4-SSC task, and for the NREM2 (majority) class for the 5-SSC task;
- The **ADA25** is the best (coupled with the ensemble) in the NREM3 (minority) class for the 4-SSC task.

Hybrid scenarios, where we coupled oversampling with CSL, are preferable if we consider the other performance metrics (GM and F1), which can be screened in the appendix tables.

The ensembles' configuration set by the Bayesian optimization is usually with bagged trees. Only the following other methods have been fitted for some scenarios:

- **Adaptive Logistic Boosting** is the optimal ensemble method for the SW task—base scenario;
- **Adaptive Boosting** is the optimal ensemble method for the 3-SSC task—base_CSL and ADA100_CSL, 4-SSC task—ADA25, and 5-SSC task—ADA50 scenarios;
- **Gentle Adaptive Boosting** is the optimal ensemble method for the SW task—ADA100, ADA50, ADA25, ADA100_CSL, ADA50_CSL, and ADA25_CSL;
- **Random Undersampling Boosting** is the optimal ensemble method for the 3-SSC task—ADA50_CSL and ADA25_CSL, 4-SSC task—ADA50_CSL, and 5-SSC task—ADA100_CSL and ADA50_CSL.

In order to provide additional performance metrics for the best modeling scenarios outlined, we present the ROC graphs for the best model in each of the 32 experimental scenarios in Appendix B (Figures A1 and A2).

To further support the comparison among classifiers based on the global performance metrics reported in Appendix A, we conducted a Wilcoxon signed-rank test across the 32 experimental scenarios. The results of this statistical analysis are summarized in Table 8, which reports the p -values for the pairwise comparisons between the ensemble model and the other evaluated classifiers (ANN, DT, and KNN), considering both macro accuracy and IAM. As shown in Table 8, the ensemble model achieved statistically significant improvements over the other models in all comparisons (all p -values < 0.001).

Table 8. Wilcoxon signed-rank test results for the comparison between the ensemble model and the other evaluated classifiers across the 32 experimental scenarios, using macro accuracy and IAM as global performance metrics.

Comparison	ACCmacro p -Value	IAM p -Value
ENSopt vs. ANNOpt	7.95×10^{-7}	2.50×10^{-7}
ENSopt vs. DTopt	7.94×10^{-7}	6.03×10^{-6}
ENSopt vs. KNNopt	1.16×10^{-8}	4.40×10^{-5}

4.5. Model Explainability Analysis

To provide additional insights into model behavior and support the interpretability of the proposed solutions, a global SHAP (SHapley Additive exPlanations) analysis was conducted on the overall best-performing models identified for each SSC task.

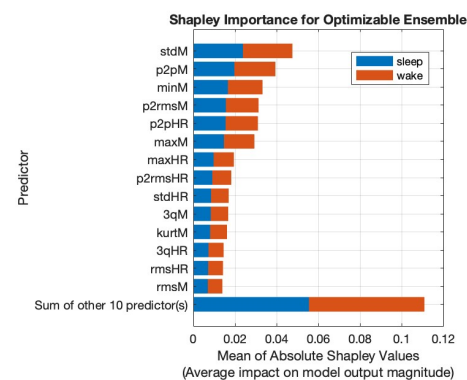
Global SHAP values were computed on a representative subset of the test set (500 samples) using the interventional approach. Shapley values were estimated via Monte Carlo sampling (500 samples) with a maximum of 128 feature subsets. Feature importance was quantified as the mean absolute SHAP value aggregated across all classes, enabling a global interpretation of model behavior.

Regarding model selection, for the 2SSC task, the optimized ensemble model under the base_CSL scenario clearly provides the best overall performance. For the 3SSC task, no single model dominates across all classes; however, the optimized ensemble under the ADA100 scenario offers the most balanced performance, particularly for NREM and wake stages, while the KNN model performs better for REM. Considering the typical trade-off between majority and minority classes, the SHAP analysis is conducted on the optimized ensemble under the ADA100 scenario. For the 4SSC task, the optimized ensemble under the base_CSL scenario achieves the best or near-best performance across all classes, particularly for REM and wake stages, while remaining competitive for NREM12 and NREM3. Similarly,

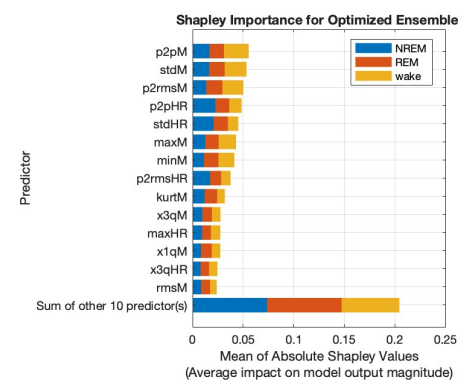
for the 5SSC task, the optimized ensemble under the base_CSL scenario represents the most reliable overall solution. Although it is not the top-performing model for REM and NREM1 (where KNN performs better), it provides superior or near-optimal performance for the remaining classes, making it the most balanced choice.

The resulting SHAP-based feature importance analyses for each SSC task are presented below.

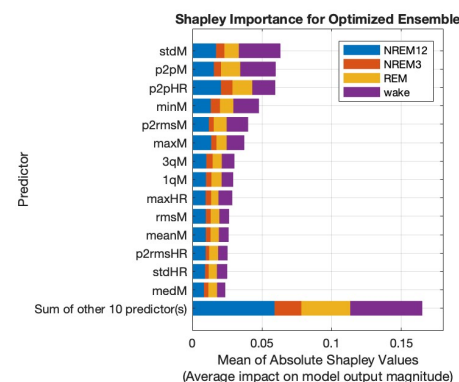
For the 2SSC task (Figure 10a), the SHAP analysis shows that model predictions are mainly driven by variability- and range-based features derived from both motion (e.g., stdM, p2pM, rmsM) and heart rate (e.g., stdHR, p2pHR). Mean-based descriptors are not among the most relevant predictors, indicating that the discrimination between sleep and wake states primarily relies on dynamic fluctuations rather than on average signal levels.



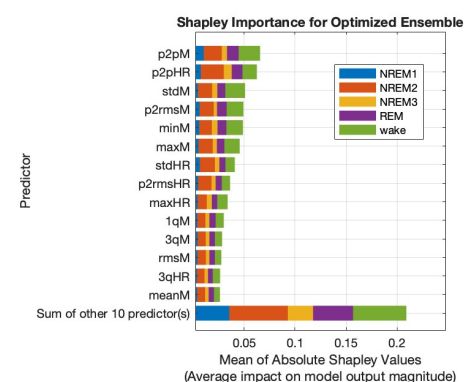
(a) 2SSC—optimized ensemble model under base_CSL scenario.



(b) 3SSC—optimized ensemble model under ADA100 scenario.



(c) 4SSC—optimized ensemble model under base_CSL scenario.



(d) 5SSC—optimized ensemble model under base_CSL scenario.

Figure 10. Global SHAP-based feature importance (top 14) for each of the four SSC tasks.

For the 3SSC task (Figure 10b), the same variability-driven behavior is observed, with both motion and heart rate features contributing significantly. However, distributional descriptors (e.g., quartiles and kurtosis) begin to emerge among the relevant predictors, and feature importance becomes more evenly distributed, reflecting the increased complexity of multi-stage classification.

For the 4SSC task (Figure 10c), variability- and range-based features remain dominant, while distributional indicators (e.g., quartiles and median-related statistics) play a more structured role in capturing finer differences among sleep stages. In this setting, feature contributions are not uniformly shared across classes: some sleep stages (e.g., wake or NREM2) are strongly associated with the most influential predictors, whereas others are characterized by more diffuse and less distinctive patterns.

For the 5SSC task (Figure 10d), this trend is further emphasized. Although variability-driven features still provide the largest contributions, feature importance is more widely distributed and increasingly class-dependent. Certain stages are clearly associated with the dominant predictors, while others are harder to distinguish and rely on a combination of lower-impact features, confirming the higher intrinsic difficulty of the task.

Overall, the SHAP analysis reveals a consistent progression across SSC tasks: variability-based features dominate in simpler settings, while distributional descriptors and a more distributed importance structure become increasingly relevant as task complexity grows. These findings suggest that reliable sleep stage classification with minimal and non-invasive signals is feasible, provided that their dynamic and statistical properties are adequately exploited, in line with the physiological complexity of sleep. This behavior is evident from the class-wise SHAP contributions, where dominant features show unbalanced relevance across sleep stages.

5. Conclusions

This study contributes to the field of data analytics and cognitive computing by investigating the feasibility of sleep stage classification (SSC), which can be achieved using a minimal and non-invasive set of physiological signals—heart rate and motion—combined with imbalance-aware machine learning strategies (RQ1). Through a systematic evaluation of 32 experimental scenarios across multiple classification granularities and model families on the PhysioNet DREAMT dataset, we show that ensemble-based approaches provide robust and consistent performance even under severe class imbalance conditions.

In particular, the ensemble model generally outperformed the other considered classifiers across most of the evaluated scenarios. This confirms the ensemble's superior capability in handling complex and imbalanced multiclass settings.

Our experiments also reveal that the effectiveness of imbalance management strategies depends on the intensity of oversampling and its interaction with algorithm-level techniques. Moderate configurations, such as ADA25, and baseline configurations combined with cost-sensitive learning (base_CSL), tended to provide better performance across several tasks. Conversely, aggressive oversampling strategies (e.g., ADA100) did not consistently improve classification performance and, in some cases, led to reduced effectiveness. This suggests that excessive synthetic sample generation might negatively affect model generalization.

From a data-driven perspective, these results highlight the effectiveness of combining data-level and algorithm-level imbalance management techniques to enhance model reliability in real-world scenarios characterized by skewed class distributions (RQ2). The observed stability of ensemble models across different tasks suggests their potential suitability for cognitive computing applications, where decision robustness and generalization are critical requirements.

A further key insight emerges from the SHAP-based explainability analysis, which reveals a progressive shift in feature relevance across SSC tasks: while variability- and range-based descriptors dominate simpler settings, more complex classifications increasingly rely on distributional features and a more distributed, class-dependent contribution of predictors. This behavior reflects the intrinsic physiological complexity of sleep and highlights the importance of capturing subtle signal dynamics when moving towards finer-grained sleep stage discrimination.

Beyond methodological aspects, the proposed framework supports scalable and computationally efficient analytics for non-invasive sleep monitoring. The reliance on easily collectible signals enables long-term, continuous data acquisition and integration into large-scale monitoring pipelines (RQ1), aligning with big data paradigms and distributed

cognitive systems. Such characteristics are particularly relevant for applications requiring adaptive and personalized insights derived from heterogeneous physiological data streams.

Finally, this work emphasizes the importance of adopting imbalance-aware evaluation metrics, such as the Matthews Correlation Coefficient, to avoid misleading conclusions based solely on global accuracy. Although validated in the context of SSC, the proposed experimental framework and methodological insights are broadly applicable to other cognitive computing and biomedical data analytics tasks affected by class imbalance.

Despite the promising results, some limitations should be addressed. The results depend on the selected feature extraction pipeline and modelling configurations adopted in this study. Although multiple models and imbalance management strategies were systematically evaluated, alternative feature representations or hyperparameter settings could lead to different performance outcomes. Furthermore, the experiments were conducted on a single publicly available dataset (PhysioNet DREAMT), which may limit the generalizability of the findings to other populations, sensor configurations, or acquisition settings. However, the use of a publicly available dataset and clearly defined experimental scenarios supports their reproducibility. Future work will include further validation on additional datasets and cross-dataset evaluations to fully assess the generalization capabilities of the proposed methodology. Additionally, the proposed framework could be extended by investigating imbalance-aware training strategies within deep learning architectures, such as CNN or LSTM models applied to raw wearable signals. Future research will also investigate subject-independent evaluation protocols to further assess the generalizability of the proposed approach, particularly in the presence of highly imbalanced sleep stage distributions.

Author Contributions: Conceptualization, L.S., A.B., S.B. and S.R.; methodology, L.S., M.E. and S.R.; software, L.S. and M.E.; validation, L.S., A.B. and S.B.; formal analysis, L.S. and P.P.; investigation, L.S.; resources, P.P.; data curation, L.S.; writing—original draft preparation, L.S., S.B., M.E. and S.R.; writing—review and editing, A.B., S.B., M.E., S.R. and P.P.; visualization, L.S., M.E. and S.R.; supervision, P.P.; project administration, A.B. and P.P.; funding acquisition, P.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used for experiments presented in this article can be freely and openly accessed at Physionet: [16].

Acknowledgments: This study has been partially promoted within the project entitled “E-BED: Empowered Bed for Elderly and Disability”, funded within PR FESR 2021–2027 program—AXIS 1—SPECIFIC OBJECTIVE 1.1—ACTION 1.1.1—call “R&D TO INNOVATE THE MARCHE REGION”—Incentives for businesses to undertake industrial research and experimental development activities within the scope of the Regional Strategy for Smart Specialization, CUP B29J24000930007. During the preparation of this manuscript/study, the author(s) used ChatGPT (free version) to assist with language editing and to improve the clarity of the English expression. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Complete Results Tables

Table A1. $F1_c$ score, GM_c and MCC_c for every model and scenario in the SW task (best overall results for each metric in bold).

Class	Scenario	ANN			DT			ENS			KNN		
		F1	GM	MCC	F1	GM	MCC	F1	GM	MCC	F1	GM	MCC
sleep	ADA100	0.791	0.652	0.293	0.766	0.638	0.257	0.854	0.704	0.430	0.760	0.676	0.312
	ADA100_CSL	0.728	0.669	0.297	0.635	0.618	0.231	0.855	0.708	0.436	0.774	0.642	0.267
	ADA25	0.861	0.610	0.357	0.862	0.617	0.366	0.878	0.642	0.429	0.849	0.658	0.371
	ADA25_CSL	0.783	0.675	0.315	0.786	0.670	0.311	0.878	0.640	0.428	0.796	0.684	0.337
	ADA50	0.830	0.647	0.332	0.826	0.658	0.338	0.870	0.690	0.448	0.853	0.615	0.347
	ADA50_CSL	0.760	0.667	0.297	0.743	0.662	0.285	0.870	0.669	0.431	0.740	0.662	0.285
	base	0.877	0.576	0.392	0.877	0.558	0.385	0.887	0.618	0.455	0.872	0.542	0.354
	base_CSL	0.850	0.670	0.386	0.850	0.648	0.364	0.879	0.670	0.455	0.854	0.661	0.385
wake	ADA100	0.492	0.652	0.293	0.473	0.638	0.257	0.576	0.704	0.430	0.514	0.676	0.312
	ADA100_CSL	0.508	0.669	0.297	0.472	0.618	0.231	0.581	0.708	0.436	0.478	0.642	0.267
	ADA25	0.483	0.610	0.357	0.491	0.617	0.366	0.532	0.642	0.429	0.521	0.658	0.371
	ADA25_CSL	0.509	0.675	0.315	0.505	0.670	0.311	0.531	0.640	0.428	0.523	0.684	0.337
	ADA50	0.502	0.647	0.332	0.511	0.658	0.338	0.575	0.690	0.448	0.485	0.615	0.347
	ADA50_CSL	0.503	0.667	0.297	0.498	0.662	0.285	0.554	0.669	0.431	0.498	0.662	0.285
	base	0.466	0.576	0.392	0.448	0.558	0.385	0.522	0.618	0.455	0.425	0.542	0.354
	base_CSL	0.536	0.670	0.386	0.513	0.648	0.364	0.564	0.670	0.455	0.530	0.661	0.385

Table A2. $F1_c$ score, GM_c and MCC_c for every model and scenario in the 3-SSC task (best overall results for each metric in bold).

Class	Scenario	ANNopt			DTopt			ENSopt			KNNopt		
		F1	GM	MCC	F1	GM	MCC	F1	GM	MCC	F1	GM	MCC
NREM	ADA100	0.680	0.648	0.286	0.671	0.626	0.242	0.789	0.729	0.448	0.657	0.614	0.219
	ADA100_CSL	0.505	0.557	0.227	0.640	0.625	0.252	0.802	0.721	0.448	0.660	0.611	0.212
	ADA25	0.749	0.649	0.308	0.717	0.635	0.267	0.817	0.672	0.424	0.780	0.598	0.292
	ADA25_CSL	0.649	0.645	0.299	0.681	0.639	0.267	0.777	0.726	0.437	0.739	0.661	0.317
	ADA50	0.730	0.652	0.301	0.692	0.633	0.257	0.802	0.701	0.427	0.734	0.643	0.289
	ADA50_CSL	0.621	0.627	0.278	0.658	0.630	0.254	0.777	0.714	0.419	0.679	0.613	0.219
	base	0.798	0.536	0.303	0.796	0.502	0.283	0.820	0.607	0.403	0.815	0.697	0.444
	base_CSL	0.746	0.659	0.320	0.740	0.632	0.280	0.792	0.485	0.257	0.815	0.697	0.444
REM	ADA100	0.316	0.632	0.224	0.311	0.602	0.217	0.491	0.711	0.428	0.276	0.574	0.174
	ADA100_CSL	0.287	0.659	0.201	0.323	0.636	0.233	0.512	0.688	0.457	0.270	0.564	0.166
	ADA25	0.335	0.584	0.249	0.322	0.581	0.233	0.465	0.589	0.450	0.240	0.423	0.187
	ADA25_CSL	0.321	0.660	0.235	0.333	0.618	0.244	0.498	0.721	0.435	0.364	0.626	0.282
	ADA50	0.344	0.601	0.256	0.320	0.591	0.226	0.483	0.647	0.431	0.278	0.533	0.181
	ADA50_CSL	0.317	0.649	0.225	0.325	0.617	0.232	0.474	0.689	0.407	0.265	0.541	0.160
	base	0.105	0.244	0.128	0.023	0.109	0.045	0.378	0.492	0.428	0.603	0.767	0.554
	base_CSL	0.329	0.578	0.240	0.271	0.502	0.181	0.000	0.000	0.000	0.603	0.767	0.554
wake	ADA100	0.501	0.653	0.321	0.465	0.627	0.263	0.589	0.719	0.444	0.447	0.611	0.240
	ADA100_CSL	0.487	0.650	0.284	0.466	0.632	0.256	0.584	0.712	0.441	0.436	0.602	0.227
	ADA25	0.485	0.631	0.319	0.459	0.618	0.270	0.563	0.679	0.437	0.485	0.618	0.344
	ADA25_CSL	0.514	0.669	0.332	0.466	0.631	0.263	0.586	0.723	0.436	0.484	0.635	0.308
	ADA50	0.501	0.649	0.328	0.461	0.622	0.261	0.578	0.702	0.439	0.495	0.639	0.326
	ADA50_CSL	0.501	0.661	0.304	0.469	0.633	0.261	0.575	0.709	0.422	0.439	0.602	0.236
	base	0.478	0.593	0.378	0.454	0.571	0.362	0.534	0.635	0.442	0.512	0.632	0.387
	base_CSL	0.518	0.657	0.356	0.499	0.642	0.331	0.441	0.562	0.346	0.512	0.632	0.387

Table A3. $F1_c$ score, GM_c and MCC_c for every model and scenario in the 4-SSC task.

Class	Scenario	ANNopt			DTopt			ENSopt			KNNopt		
		F1	GM	MCC	F1	GM	MCC	F1	GM	MCC	F1	GM	MCC
NREM12	ADA100	0.662	0.631	0.256	0.653	0.632	0.260	0.761	0.725	0.441	0.604	0.586	0.174
	ADA100_CSL	0.521	0.570	0.238	0.589	0.602	0.233	0.634	0.664	0.391	0.604	0.586	0.174
	ADA25	0.705	0.626	0.258	0.696	0.645	0.285	0.800	0.678	0.428	0.724	0.573	0.214
	ADA25_CSL	0.529	0.573	0.230	0.646	0.625	0.247	0.749	0.721	0.433	0.659	0.603	0.202
	ADA50	0.680	0.640	0.273	0.668	0.633	0.261	0.783	0.716	0.440	0.652	0.621	0.238
	ADA50_CSL	0.516	0.565	0.227	0.611	0.612	0.237	0.742	0.712	0.415	0.631	0.603	0.204
	base	0.774	0.527	0.286	0.770	0.475	0.253	0.801	0.612	0.401	0.781	0.569	0.326
	base_CSL	0.703	0.644	0.286	0.699	0.639	0.275	0.797	0.687	0.433	0.704	0.626	0.258

Table A3. Cont.

Class	Scenario	ANNopt			DTopt			ENSopt			KNNopt		
		F1	GM	MCC	F1	GM	MCC	F1	GM	MCC	F1	GM	MCC
NREM3	ADA100	0.416	0.696	0.398	0.426	0.749	0.418	0.616	0.817	0.603	0.305	0.685	0.296
	ADA100_CSL	0.399	0.729	0.389	0.397	0.758	0.395	0.537	0.848	0.540	0.305	0.685	0.296
	ADA25	0.412	0.698	0.394	0.433	0.731	0.420	0.641	0.743	0.638	0.258	0.589	0.235
	ADA25_CSL	0.278	0.786	0.308	0.409	0.747	0.402	0.604	0.828	0.595	0.354	0.703	0.344
	ADA50	0.399	0.688	0.381	0.417	0.728	0.405	0.643	0.800	0.630	0.377	0.724	0.368
	ADA50_CSL	0.332	0.744	0.335	0.366	0.729	0.360	0.545	0.790	0.533	0.329	0.690	0.318
	base	0.373	0.510	0.401	0.000	0.000	0.000	0.587	0.671	0.606	0.520	0.638	0.528
REM	base_CSL	0.375	0.644	0.353	0.425	0.669	0.405	0.636	0.742	0.633	0.323	0.633	0.302
	ADA100	0.312	0.579	0.220	0.342	0.628	0.253	0.503	0.718	0.441	0.251	0.535	0.143
	ADA100_CSL	0.315	0.627	0.220	0.328	0.636	0.238	0.438	0.754	0.379	0.251	0.535	0.143
	ADA25	0.303	0.545	0.216	0.378	0.620	0.299	0.499	0.613	0.488	0.153	0.333	0.094
	ADA25_CSL	0.294	0.632	0.198	0.349	0.627	0.263	0.515	0.726	0.455	0.259	0.514	0.162
	ADA50	0.313	0.584	0.222	0.332	0.598	0.244	0.502	0.670	0.450	0.283	0.557	0.186
	ADA50_CSL	0.303	0.642	0.210	0.313	0.610	0.220	0.483	0.714	0.419	0.263	0.538	0.161
wake	base	0.085	0.216	0.114	0.004	0.045	0.046	0.360	0.474	0.421	0.214	0.361	0.237
	base_CSL	0.313	0.559	0.223	0.298	0.546	0.206	0.501	0.607	0.498	0.244	0.467	0.157
	ADA100	0.492	0.652	0.305	0.464	0.626	0.272	0.596	0.732	0.452	0.430	0.598	0.226
	ADA100_CSL	0.487	0.657	0.275	0.469	0.637	0.265	0.585	0.738	0.426	0.430	0.598	0.226
	ADA25	0.468	0.615	0.298	0.476	0.629	0.290	0.568	0.681	0.444	0.457	0.592	0.315
	ADA25_CSL	0.507	0.657	0.329	0.465	0.625	0.267	0.597	0.733	0.446	0.434	0.593	0.239
	ADA50	0.498	0.649	0.319	0.472	0.630	0.281	0.592	0.717	0.453	0.463	0.621	0.271
	ADA50_CSL	0.495	0.657	0.295	0.467	0.631	0.264	0.580	0.718	0.427	0.443	0.606	0.242
	base	0.468	0.585	0.371	0.459	0.574	0.372	0.542	0.643	0.450	0.482	0.598	0.380
	base_CSL	0.527	0.667	0.363	0.490	0.639	0.312	0.591	0.704	0.463	0.503	0.646	0.335

Table A4. F1_c score, GM_c and MCC_c for every model and scenario in the 5-SSC task.

Class	Scenario	ANNopt			DTopt			ENSopt			KNNopt		
		F1	GM	MCC	F1	GM	MCC	F1	GM	MCC	F1	GM	MCC
NREM1	ADA100	0.218	0.499	0.093	0.210	0.487	0.084	0.259	0.515	0.154	0.197	0.466	0.072
	ADA100_CSL	0.211	0.533	0.069	0.218	0.508	0.090	0.232	0.488	0.121	0.198	0.463	0.073
	ADA25	0.120	0.283	0.069	0.191	0.434	0.076	0.162	0.328	0.124	0.178	0.407	0.069
	ADA25_CSL	0.226	0.503	0.103	0.192	0.458	0.064	0.268	0.504	0.171	0.175	0.410	0.061
	ADA50	0.192	0.416	0.092	0.197	0.456	0.076	0.205	0.396	0.136	0.211	0.461	0.100
	ADA50_CSL	0.220	0.520	0.090	0.206	0.489	0.077	0.240	0.485	0.136	0.182	0.428	0.064
	base	0.004	0.045	0.016	0.023	0.109	0.048	0.094	0.229	0.120	0.294	0.535	0.203
NREM2	base_CSL	0.000	0.000	0.000	0.200	0.439	0.097	0.207	0.389	0.153	0.293	0.581	0.192
	ADA100	0.536	0.593	0.258	0.597	0.634	0.300	0.706	0.721	0.451	0.551	0.597	0.237
	ADA100_CSL	0.141	0.276	0.118	0.520	0.585	0.277	0.658	0.682	0.381	0.567	0.610	0.256
	ADA25	0.664	0.637	0.285	0.624	0.642	0.289	0.742	0.699	0.435	0.599	0.620	0.248
	ADA25_CSL	0.510	0.579	0.276	0.554	0.606	0.271	0.694	0.715	0.449	0.588	0.611	0.229
	ADA50	0.614	0.632	0.268	0.601	0.631	0.280	0.738	0.715	0.442	0.621	0.647	0.308
	ADA50_CSL	0.409	0.504	0.225	0.523	0.585	0.261	0.682	0.703	0.420	0.564	0.597	0.210
NREM3	base	0.701	0.535	0.281	0.703	0.529	0.284	0.744	0.658	0.419	0.742	0.706	0.439
	base_CSL	0.629	0.642	0.288	0.638	0.651	0.305	0.739	0.722	0.451	0.690	0.710	0.433
	ADA100	0.364	0.769	0.370	0.405	0.717	0.392	0.599	0.818	0.588	0.361	0.713	0.352
	ADA100_CSL	0.199	0.781	0.242	0.403	0.744	0.396	0.498	0.751	0.482	0.381	0.729	0.373
	ADA25	0.397	0.752	0.394	0.437	0.739	0.426	0.644	0.784	0.633	0.334	0.720	0.333
	ADA25_CSL	0.351	0.775	0.363	0.391	0.743	0.387	0.594	0.856	0.592	0.359	0.695	0.348
	ADA50	0.335	0.764	0.346	0.412	0.734	0.403	0.618	0.760	0.608	0.408	0.760	0.405
REM	ADA50_CSL	0.285	0.776	0.310	0.385	0.741	0.381	0.523	0.794	0.514	0.318	0.692	0.311
	base	0.245	0.393	0.289	0.219	0.363	0.281	0.596	0.680	0.612	0.594	0.756	0.580
	base_CSL	0.354	0.651	0.335	0.464	0.724	0.448	0.673	0.790	0.665	0.470	0.842	0.483
	ADA100	0.306	0.605	0.210	0.342	0.597	0.257	0.508	0.719	0.447	0.277	0.544	0.178
	ADA100_CSL	0.259	0.596	0.150	0.332	0.614	0.242	0.422	0.654	0.350	0.294	0.560	0.200
	ADA25	0.282	0.525	0.190	0.352	0.593	0.269	0.502	0.643	0.464	0.282	0.539	0.187
	ADA25_CSL	0.325	0.640	0.233	0.336	0.604	0.247	0.505	0.740	0.443	0.277	0.528	0.182
	ADA50	0.283	0.548	0.187	0.356	0.600	0.274	0.514	0.666	0.468	0.332	0.593	0.244
	ADA50_CSL	0.292	0.610	0.193	0.330	0.609	0.239	0.484	0.714	0.420	0.253	0.509	0.153
	base	0.064	0.186	0.086	0.031	0.126	0.065	0.425	0.537	0.446	0.628	0.757	0.588
	base_CSL	0.322	0.606	0.228	0.295	0.541	0.202	0.524	0.657	0.488	0.579	0.808	0.531

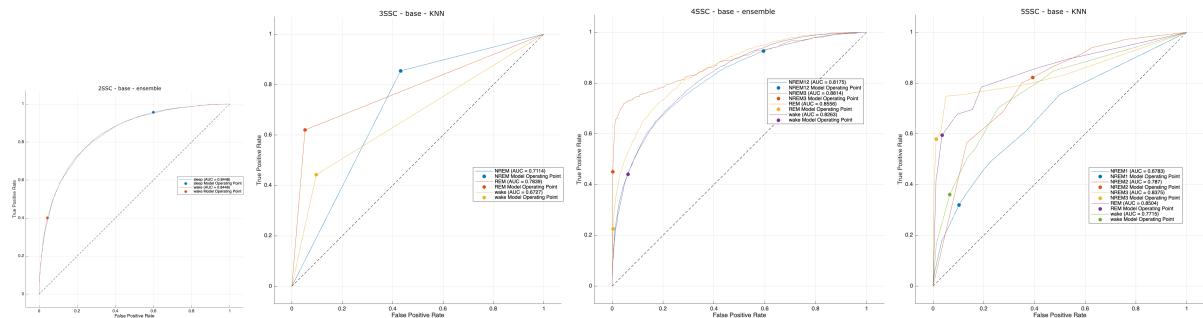
Table A4. Cont.

Class	Scenario	ANNopt			DTopt			ENSopt			KNNopt		
		F1	GM	MCC	F1	GM	MCC	F1	GM	MCC	F1	GM	MCC
wake	ADA100	0.457	0.598	0.307	0.442	0.595	0.266	0.572	0.698	0.433	0.425	0.583	0.240
	ADA100_CSL	0.380	0.510	0.295	0.438	0.599	0.245	0.534	0.675	0.375	0.435	0.591	0.254
	ADA25	0.504	0.632	0.367	0.454	0.603	0.282	0.582	0.700	0.451	0.446	0.596	0.275
	ADA25_CSL	0.472	0.617	0.308	0.450	0.607	0.260	0.586	0.715	0.441	0.427	0.584	0.243
	ADA50	0.485	0.618	0.341	0.439	0.594	0.256	0.577	0.701	0.438	0.449	0.598	0.277
	ADA50_CSL	0.461	0.604	0.299	0.438	0.597	0.244	0.557	0.691	0.405	0.410	0.568	0.224
	base	0.498	0.621	0.370	0.494	0.616	0.369	0.583	0.695	0.458	0.464	0.581	0.369
	base_CSL	0.524	0.675	0.342	0.494	0.636	0.329	0.599	0.728	0.453	0.485	0.604	0.374

Table A5. Global metrics (IAM and macro accuracy) of every model under every modeling scenario for the different SSC tasks.

Task	Scenario	ANNopt		DTopt		ENSopt		KNNopt	
		IAM	ACCmacro	IAM	ACCmacro	IAM	ACCmacro	IAM	ACCmacro
SW	ADA100	0.186	0.704	0.113	0.676	0.421	0.783	0.098	0.679
	ADA100_CSL	0.022	0.649	−0.154	0.569	0.424	0.784	0.136	0.685
	ADA25	0.236	0.781	0.246	0.783	0.280	0.806	0.344	0.770
	ADA25_CSL	0.148	0.698	0.158	0.701	0.276	0.806	0.188	0.714
	ADA50	0.327	0.746	0.308	0.743	0.385	0.800	0.246	0.771
	ADA50_CSL	0.097	0.676	0.055	0.660	0.338	0.798	0.048	0.658
	base	0.163	0.800	0.136	0.799	0.228	0.817	0.114	0.791
	base_CSL	0.367	0.773	0.320	0.770	0.333	0.811	0.342	0.777
3-SSC	ADA100	−0.128	0.574	−0.152	0.562	0.179	0.702	−0.195	0.542
	ADA100_CSL	−0.351	0.448	−0.206	0.538	0.249	0.717	−0.197	0.541
	ADA25	0.008	0.636	−0.063	0.602	0.092	0.731	−0.102	0.670
	ADA25_CSL	−0.179	0.552	−0.137	0.572	0.146	0.692	−0.005	0.627
	ADA50	−0.007	0.622	−0.111	0.579	0.195	0.715	−0.032	0.618
	ADA50_CSL	−0.222	0.530	−0.170	0.554	0.157	0.688	−0.158	0.558
	base	−0.238	0.692	−0.297	0.687	−0.062	0.731	0.204	0.725
	base_CSL	0.025	0.638	−0.002	0.628	−0.315	0.681	0.204	0.725
4-SSC	ADA100	−0.168	0.560	−0.191	0.548	0.144	0.681	−0.327	0.491
	ADA100_CSL	−0.340	0.471	−0.282	0.504	−0.137	0.582	−0.327	0.491
	ADA25	−0.112	0.590	−0.091	0.590	0.098	0.718	−0.306	0.599
	ADA25_CSL	−0.369	0.459	−0.196	0.545	0.120	0.674	−0.225	0.538
	ADA50	−0.147	0.572	−0.163	0.562	0.237	0.702	−0.225	0.539
	ADA50_CSL	−0.372	0.459	−0.273	0.514	0.064	0.659	−0.278	0.516
	base	−0.328	0.663	−0.496	0.656	−0.088	0.713	−0.194	0.676
	base_CSL	−0.093	0.599	−0.089	0.587	0.118	0.719	−0.150	0.589
5-SSC	ADA100	−0.397	0.424	−0.304	0.461	−0.020	0.588	−0.380	0.423
	ADA100_CSL	−0.684	0.230	−0.371	0.417	−0.142	0.536	−0.355	0.437
	ADA25	−0.314	0.529	−0.250	0.489	−0.048	0.635	−0.343	0.460
	ADA25_CSL	−0.401	0.417	−0.346	0.437	−0.048	0.587	−0.339	0.450
	ADA50	−0.328	0.479	−0.286	0.468	−0.012	0.626	−0.290	0.481
	ADA50_CSL	−0.494	0.359	−0.375	0.417	−0.096	0.565	−0.395	0.426
	base	−0.525	0.573	−0.540	0.573	−0.207	0.640	−0.007	0.619
	base_CSL	−0.346	0.511	−0.217	0.505	0.000	0.636	−0.164	0.563

Appendix B. Roc-Auc Graphs

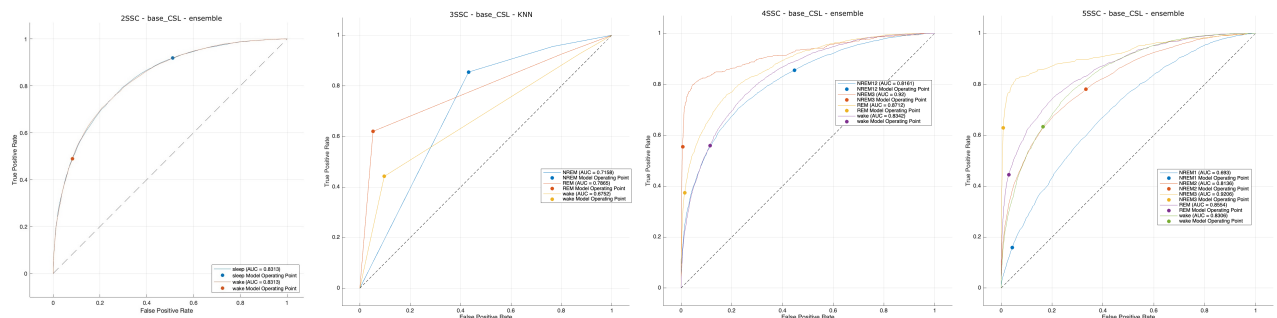


(a) Base - 2SSC - Opt. Ens

(b) Base - 3SSC - kNN

(c) Base - 4SSC - Opt. Ens

(d) Base - 5SSC - kNN

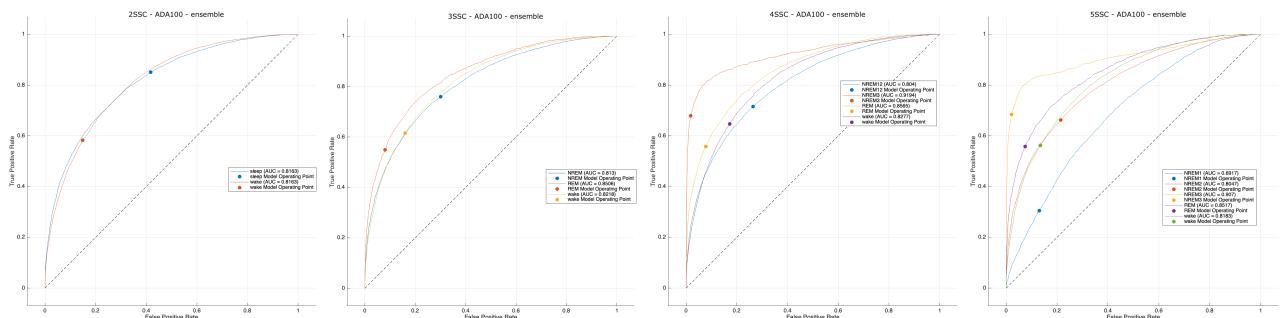


(e) Base_CSL - 2SSC - Opt. Ens

(f) Base_CSL - 3SSC - kNN

(g) Base_CSL - 4SSC - Opt. Ens

(h) Base_CSL - 5SSC - Opt. Ens

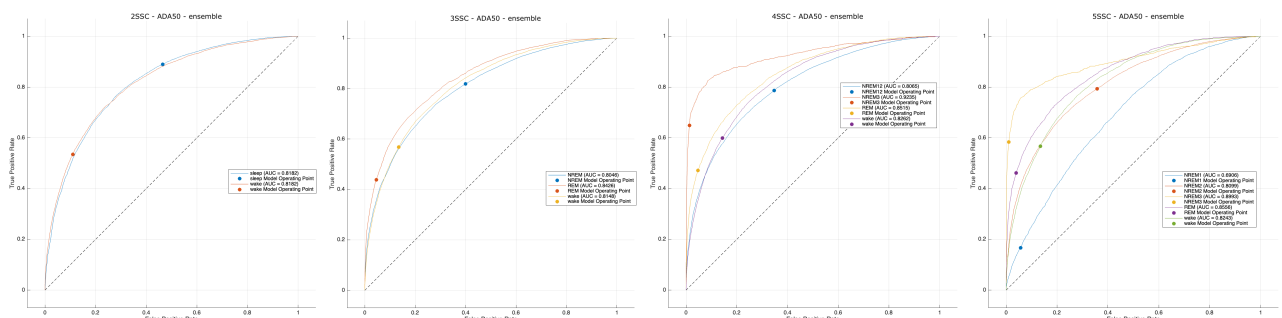


(i) ADA100 - 2SSC - Opt. Ens

(j) ADA100 - 3SSC - Opt. Ens

(k) ADA100 - 4SSC - Opt. Ens

(l) ADA100 - 5SSC - Opt. Ens



(m) ADA50 - 2SSC - Opt. Ens

(n) ADA50 - 3SSC - Opt. Ens

(o) ADA50 - 4SSC - Opt. Ens

(p) ADA50 - 5SSC - Opt. Ens

Figure A1. Cont.

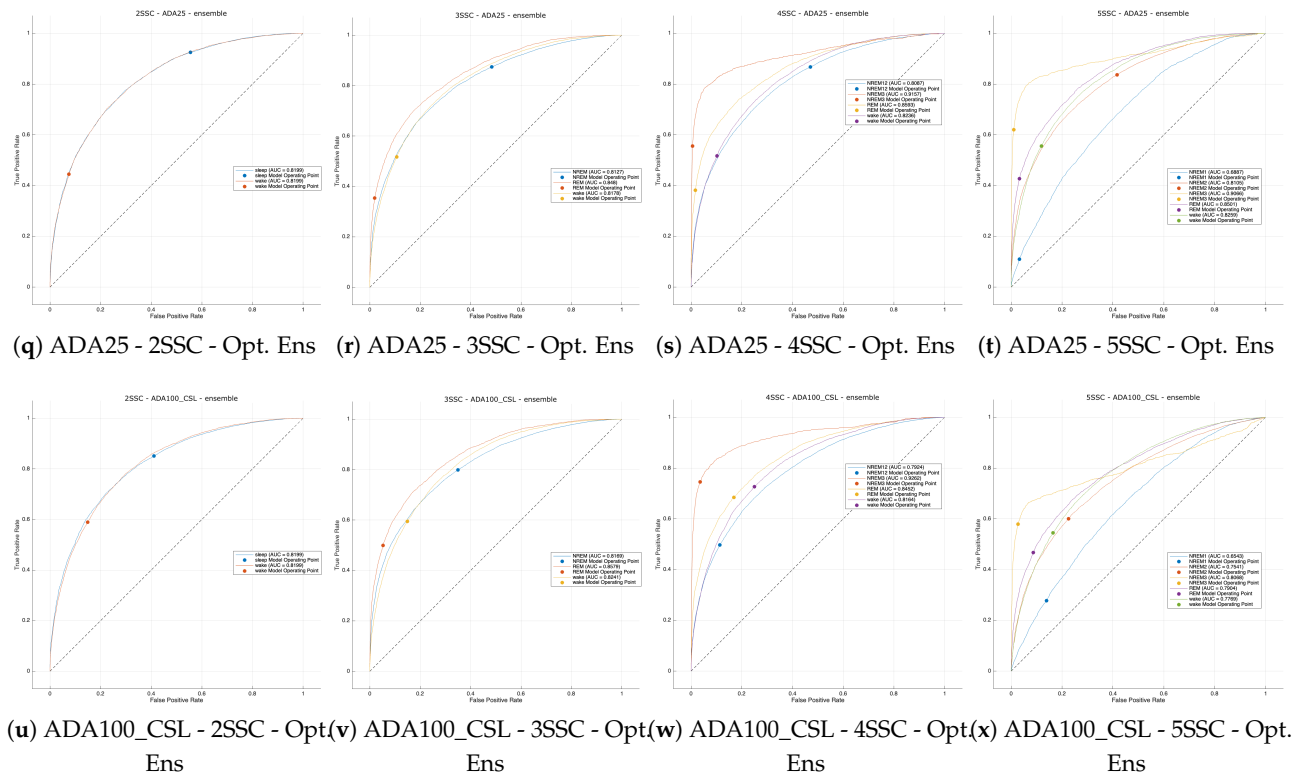


Figure A1. ROC-AUC curves for each scenario - task - best model, under the base, base_CSL, ADA100, ADA50, ADA25, and ADA100_CSL scenarios (different scenarios in different rows, different SSC tasks in each column, as described in sub-captions).

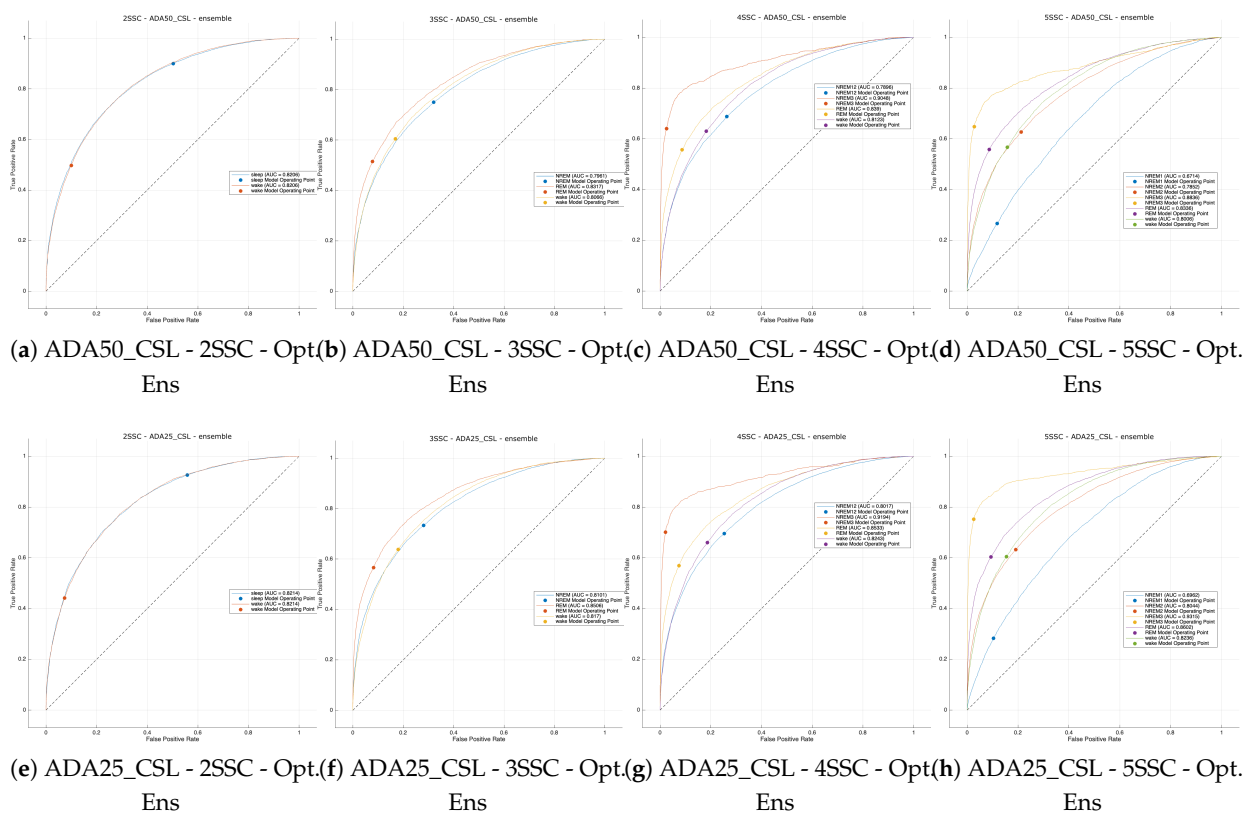


Figure A2. ROC-AUC curves for each scenario - task - best model, under the ADA50_CSL, and ADA25_CSL scenarios.

References

- Hussain, Z.; Sheng, Q.Z.; Zhang, W.E.; Ortiz, J.; Pouriyeh, S. Non-invasive Techniques for Monitoring Different Aspects of Sleep: A comprehensive review. *ACM Trans. Comput. Healthc.* **2022**, *3*, 1–26. [\[CrossRef\]](#)
- Boostani, R.; Karimzadeh, F.; Nami, M. A comparative review on sleep stage classification methods in patients and healthy individuals. *Comput. Methods Programs Biomed.* **2017**, *140*, 77–91. [\[CrossRef\]](#)
- Kryger, M.H.; Roth, T.; Dement, W.C. *Principles and Practice of Sleep Medicine E-Book: Expert Consult-Online and Print*; Elsevier Health Science: Amsterdam, The Netherlands, 2010.
- Zhou, D.; Xu, Q.; Wang, J.; Xu, H.; Kettunen, L.; Chang, Z.; Cong, F. Alleviating Class Imbalance Problem in Automatic Sleep Stage Classification. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 406612. [\[CrossRef\]](#)
- Chinoy, E.D.; Cuellar, J.A.; Huwa, K.E.; Jameson, J.T.; Watson, C.H.; Bessman, S.C.; Hirsch, D.A.; Cooper, A.D.; Drummond, S.P.; Markwald, R.R. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep* **2021**, *44*, zsaa291. [\[CrossRef\]](#)
- Tal, A.; Shinar, Z.; Shaki, D.; Codish, S.; Goldbart, A. Validation of contact-free sleep monitoring device with comparison to polysomnography. *J. Clin. Sleep Med.* **2017**, *13*, 517–522. [\[CrossRef\]](#)
- Gaiduk, M.; Penzel, T.; Ortega, J.A.; Seepold, R. Automatic sleep stages classification using respiratory, heart rate and movement signals. *Physiol. Meas.* **2018**, *39*, 124008. [\[CrossRef\]](#)
- Morokuma, S.; Hayashi, T.; Kanegae, M.; Mizukami, Y.; Asano, S.; Kimura, I.; Tateizumi, Y.; Ueno, H.; Ikeda, S.; Niizeki, K. Deep learning-based sleep stage classification with cardiorespiratory and body movement activities in individuals with suspected sleep disorders. *Sci. Rep.* **2023**, *13*, 17730. [\[CrossRef\]](#)
- Mitsukura, Y.; Fukunaga, K.; Yasui, M.; Mimura, M. Sleep stage detection using only heart rate. *Health Inform. J.* **2020**, *26*, 376–387. [\[CrossRef\]](#) [\[PubMed\]](#)
- Brunner, C.; Hofer, F. SleepECG: A Python package for sleep staging based on heart rate. *J. Open Source Softw.* **2023**, *8*, 5411. [\[CrossRef\]](#)
- Yi, R.; Enayati, M.; Keller, J.M.; Popescu, M.; Skubic, M. Non-invasive in-home sleep stage classification using a ballistocardiography bed sensor. In *Proceedings of the 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Chicago, IL, USA, 19–22 May 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–4.
- Pierleoni, P.; Belli, A.; Palma, L.; Pernini, L.; Valenti, S. An accurate device for real-time altitude estimation using data fusion algorithms. In *Proceedings of the 2014 IEEE/ASME 10th International Conference on Mechatronic and Embedded Systems and Applications (MESA), Senigallia, Italy, 10–12 September 2014*; pp. 1–5. [\[CrossRef\]](#)
- Pardamean, B.; Budiarto, A.; Mahesworo, B.; Hidayat, A.A.; Sudigyo, D. Supervised Learning for Imbalance Sleep Stage Classification Problem. *Commun. Math. Biol. Neurosci.* **2023**, *2023*, 131. [\[CrossRef\]](#)
- Liang, Z.; Chapa-Martell, M.A. A Multi-Level Classification Approach for Sleep Stage Prediction With Processed Data Derived From Consumer Wearable Activity Trackers. *Front. Digit. Health* **2021**, *3*, 665946. [\[CrossRef\]](#) [\[PubMed\]](#)
- Liang, Z.; Chapa-Martell, M.A. Combining Resampling and Machine Learning to Improve Sleep-Wake Detection of Fitbit Wristbands. In *Proceedings of the 2019 IEEE International Conference on Healthcare Informatics (ICHI), Xi'an, China, 10–13 June 2019*; pp. 1–3. [\[CrossRef\]](#)
- Wang, K.; Yang, J.; Shetty, A.; Dunn, J. DREAMT: Dataset for Real-time sleep stage Estimation using Multisensor wearable Technology. *PhysioNet* **2024**, RRID:SCR_007345. [\[CrossRef\]](#)
- Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **2000**, *101*, e215–e220. [\[CrossRef\]](#) [\[PubMed\]](#)
- Logacjov, A.; Bach, K.; Mork, P.J. Long-term self-supervised learning for accelerometer-based sleep–wake recognition. *Eng. Appl. Artif. Intell.* **2025**, *141*, 109758. [\[CrossRef\]](#)
- Ohayon, M.; Wickwire, E.; Hirshkowitz, M.; Albert, S.; Avidan, A.; Daly, F.; Dauvilliers, Y.; Ferri, R.; Fung, C.; Gozal, D.; et al. National Sleep Foundation's sleep quality recommendations: First report. *Sleep Health* **2017**, *3*, 6–19. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wang, L.; Han, M.; Li, X.; Zhang, N.; Cheng, H. Review of Classification Methods on Unbalanced Data Sets. *IEEE Access* **2021**, *9*, 64606–64628. [\[CrossRef\]](#)
- Kaur, H.; Pannu, H.S.; Malhi, A.K. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Comput. Surv.* **2019**, *52*, 79. [\[CrossRef\]](#)
- Jafarigol, E.; Trafalis, T. A Review of Machine Learning Techniques in Imbalanced Data and Future Trends. *arXiv* **2023**, arXiv:2310.07917. [\[CrossRef\]](#)

23. Yang, Y.; Khorshidi, H.A.; Aickelin, U. A review on over-sampling techniques in classification of multi-class imbalanced datasets: Insights for medical problems. *Front. Digit. Health* **2024**, *6*, 1430245. [\[CrossRef\]](#)
24. Altalhan, M.; Algarni, A.; Alouane, M.T.H. Imbalanced Data Problem in Machine Learning: A Review. *IEEE Access* **2025**, *13*, 13686–13699. [\[CrossRef\]](#)
25. Jadhav, A.; Mostafa, S.M.; Elmannai, H.; Karim, F.K. An Empirical Assessment of Performance of Data Balancing Techniques in Classification Task. *Appl. Sci.* **2022**, *12*, 3928. [\[CrossRef\]](#)
26. Balakrishnan, A.; Medikonda, J.; Namboothiri, P.K.; Natarajan, M. Parkinson's Disease Stage Classification with Gait Analysis using Machine Learning Techniques and SMOTE-based Approach for Class Imbalance Problem. In *Proceedings of the 2022 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics, DISCOVER 2022—Proceedings, Shivamogga, India, 14–15 October 2022*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2022; pp. 277–281. [\[CrossRef\]](#)
27. Spelman, V.S.; Porkodi, R. A Review on Handling Imbalanced Data. In *Proceedings of the 2018 IEEE International Conference on Current Trends Toward Converging Technologies, Coimbatore, India, 1–3 March 2018*; Institute of Electrical and Electronics Engineers: Piscataway, NJ, USA, 2018; p. 115. [\[CrossRef\]](#)
28. Brandt, J.; Lanzén, E. A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification. Master's Thesis, Uppsala University, Uppsala, Sweden, 2020. Available online: <https://www.diva-portal.org/smash/get/diva2:1519153/FULLTEXT01.pdf> (accessed on 1 January 2026).
29. Abdi, L.; Hashemi, S. To Combat Multi-Class Imbalanced Problems by Means of Over-Sampling Techniques. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 238–251. [\[CrossRef\]](#)
30. Moghadas-Dastjerdi, H.; Sha-E-Tallat, H.R.; Sannachi, L.; Sadeghi-Naini, A.; Czarnota, G.J. A priori prediction of tumour response to neoadjuvant chemotherapy in breast cancer patients using quantitative CT and machine learning. *Sci. Rep.* **2020**, *10*, 10936. [\[CrossRef\]](#)
31. Mienye, I.D.; Sun, Y. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Inform. Med. Unlocked* **2021**, *25*, 10069. [\[CrossRef\]](#)
32. Araf, I.; Idri, A.; Chairi, I. Cost-sensitive learning for imbalanced medical data: A review. *Artif. Intell. Rev.* **2024**, *57*, 80. [\[CrossRef\]](#)
33. Ren, Z.; Zhu, Y.; Kang, W.; Fu, H.; Niu, Q.; Gao, D.; Yan, K.; Hong, J. Adaptive cost-sensitive learning: Improving the convergence of intelligent diagnosis models under imbalanced data. *Knowl.-Based Syst.* **2022**, *241*, 108296. [\[CrossRef\]](#)
34. Le, T.; Vo, M.T.; Vo, B.; Lee, M.Y.; Baik, S.W. A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction. *Complexity* **2019**, *2019*, 8460934. [\[CrossRef\]](#)
35. El-Amir, S.; El-Henawy, I. An Improved Model Using Oversampling Technique and Cost-Sensitive Learning for Imbalanced Data Problem. *Inf. Sci. Appl.* **2024**, *2*, 33–50. [\[CrossRef\]](#)
36. Naghavi, N.; Miller, A.; Wade, E. Towards real-time prediction of freezing of gait in patients with parkinson's disease: Addressing the class imbalance problem. *Sensors* **2019**, *19*, 3898. [\[CrossRef\]](#)
37. Pes, B. Learning from high-dimensional biomedical datasets: The issue of class imbalance. *IEEE Access* **2020**, *8*, 13527–13540. [\[CrossRef\]](#)
38. Feng, F.; Li, K.C.; Shen, J.; Zhou, Q.; Yang, X. Using Cost-Sensitive Learning and Feature Selection Algorithms to Improve the Performance of Imbalanced Classification. *IEEE Access* **2020**, *8*, 69979–69996. [\[CrossRef\]](#)
39. Solanki, Y.S.; Chakrabarti, P.; Jasinski, M.; Leonowicz, Z.; Bolshev, V.; Vinogradov, A.; Jasinska, E.; Gono, R.; Nami, M. A hybrid supervised machine learning classifier system for breast cancer prognosis using feature selection and data imbalance handling approaches. *Electronics* **2021**, *10*, 699. [\[CrossRef\]](#)
40. Ghosh, K.; Bellinger, C.; Corizzo, R.; Branco, P.; Krawczyk, B.; Japkowicz, N. The class imbalance problem in deep learning. *Mach. Learn.* **2024**, *113*, 4845–4901. [\[CrossRef\]](#)
41. Apu, K.U.; Ali, M. A Systematic Literature Review on AI Approaches to Address Data Imbalance In Machine Learning. *Front. Appl. Eng. Technol.* **2025**, *2*, 58–77. [\[CrossRef\]](#)
42. Zhu, J.; Pu, S.; He, J.; Su, D.; Cai, W.; Xu, X.; Liu, H. Processing imbalanced medical data at the data level with assisted-reproduction data as an example. *BioData Min.* **2024**, *17*, 29. [\[CrossRef\]](#) [\[PubMed\]](#)
43. McCarthy, C.; Pradhan, N.; Redpath, C.; Adler, A. Validation of the Empatica E4 wristband. In *Proceedings of the 2016 IEEE EMBS International Student Conference (ISC), Ottawa, ON, Canada, 29–31 May 2016*; pp. 1–4. [\[CrossRef\]](#)
44. Prieto, M.D.; Cirrincione, G.; Espinosa, A.G.; Ortega, J.A.; Henao, H. Bearing Fault Detection by a Novel Condition-Monitoring Scheme Based on Statistical-Time Features and Neural Networks. *IEEE Trans. Ind. Electron.* **2013**, *60*, 3398–3407. [\[CrossRef\]](#)
45. Nayana, B.R.; Geethanjali, P. Analysis of Statistical Time-Domain Features Effectiveness in Identification of Bearing Faults From Vibration Signal. *IEEE Sens. J.* **2017**, *17*, 5618–5625. [\[CrossRef\]](#)
46. Grandini, M.; Bagli, E.; Visani, G. Metrics for Multi-Class Classification: An Overview. *arXiv* 2020, arXiv:2008.05756. [\[CrossRef\]](#)

47. Chicco, D.; Jurman, G. The Matthews correlation coefficient (MCC) should replace the ROC-AUC as the standard metric for assessing binary classification. *BioData Min.* **2023**, *16*, 4. [[CrossRef](#)] [[PubMed](#)]
48. Mortaz, E. Imbalance accuracy metric for model selection in multi-class imbalance classification problems. *Knowl.-Based Syst.* **2020**, *210*, 106490. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.