International Journal of Basic and Applied Sciences, 14 (8) (2025) 174-197

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/d7tqy474
Research paper

Optimization of The Process of Identifying Sleep-Disordered Breathing Based on CNN and LSTM Recurrent Neural Networks Using PSG EEG Signals

Manoj K. *, Athiamaan R.

College of Engineering, Anna University, Chennai, India
**Corresponding author E-mail: manojkkmj@gmail.com*

Received: October 14, 2025, Accepted: December 4, 2025, Published: December 10, 2025

Abstract

One of the challenges of Clinical polysomnographic (PSG) Datasets is the volume of information they contain, as they can consist of hundreds to thousands of PSG records, and each PSG record contains more than a dozen clinical time series about eight hours in length. Manual analysis of such datasets is a slow and laborious process, which is highly dependent on the experience and skill of the sleep technologist and consequently limits PSG-based sleep-related studies. The main objective of this project is the automatic identification of arousals due to respiratory events, specifically RERA (Respiratory Effort-Related Arousal) events and events associated with apnea/hypopnea using PSG electroencephalography (EEG) signals. This paper implements technique to deal with data imbalance and improve systems performance in identifying Respiratory events, design and implements different classification systems for identifying arousals related to respiratory events using PSG EEG signals. Also analyze the performance obtained for each of the implemented classifiers compared to the models presented in the literature. For this four systems were developed based on convolutional neural networks (CNN) and Long-Short Term Memory (LSTM) recurrent neural networks. In this context, this research aims to contribute to the optimization of the process of identifying sleep-disordered breathing.

Keywords: *Respiratory Effort-Related Arousal; Polysomnographic; Electroencephalography; Long-Short Term Memory; Convolutional Neural Networks.*

1. Introduction

According to the World Health Organization (WHO), approximately 40% of the global population experiences poor sleep due to a sleep disorder. Sleep quality is closely linked to an individual's mental, physical, and psychological well-being, and inadequate sleep can lead to various health problems, including memory and learning difficulties, obesity, hypotension, and cardiovascular dysfunction (Salari et al., 2022; Devnani & Hegde, 2015). There are roughly 90 recognized sleep disorders, which are categorized into insomnia, respiratory disorders, sleep-related movement disorders, hypersomnias of central origin, parasomnias, and circadian rhythm disorders (Hospitals et al., 2018). Sleep disorders are diagnosed based on clinical and neurophysiological criteria, allowing the identification of primary causes, which may include i) dysfunction of the sleep control mechanisms, or ii) impaired function of an organ, such as the upper airway or lungs (French & Muthusamy, 2016). One of the most commonly used diagnostic tools is clinical polysomnography (PSG) (French & Muthusamy, 2016). Despite the prevalence of sleep disorders, many individuals remain undiagnosed or misdiagnosed, highlighting the need for efficient and accurate assessment methods (Dauvilliers et al., 2020). Sleep evaluation offers an opportunity to detect conditions such as apnea, hypopnea, RERA events, bruxism, and airway obstruction through clinical observation and sleep testing (Korompili et al., 2021).

Most prior work in this area has focused on either a single disease or specific parameters, such as sleep stages, creating an opportunity for research into detecting sleep arousals based solely on EEG recordings. EEG-based approaches can reduce the complexity and cost of diagnostic evaluation, while also improving comfort in pediatric populations, as fewer sensors are required (Salari et al., 2022). Furthermore, since epileptic changes in children predominantly occur during sleep, identifying arousals exclusively through EEG recordings provides additional diagnostic value (Devnani & Hegde, 2015).

This research focuses on detecting awakenings (arousals) associated with three specific respiratory disorders: apnea, hypopnea, and respiratory effort-related arousals (RERA). Sleep arousals can occur spontaneously, during transitions between sleep states, or in association with other sleep disorders or environmental stimuli (Hospitals et al., 2018). According to the American Academy of Sleep Medicine, arousals are characterized by abrupt changes in EEG frequency (French & Muthusamy, 2016). However, certain physiological signals may also show characteristic changes depending on the sleep disorder. For instance, apnea, hypopnea, and RERA events are associated with abrupt changes in EEG waveforms, airflow, and oxygen saturation (SaO_2) (Korompili et al., 2021). Therefore, analyzing EEG together with other PSG-recorded biosignals, such as electrooculography (EOG) and electrocardiography (ECG), provides crucial information for arousal detection (Hospitals et al., 2018).

Given these considerations, this study proposes the use of deep learning neural networks (DLNNs) for automatic detection of arousals related to respiratory disorders, using PSG recordings from the PhysioNet/Computing in Cardiology (CinC) Challenge 2018 database (Hospitals et al., 2018).

2. Literature Review

EEG-based sleep and arousal analysis has undergone rapid transformations in hardware burden reduction, deeper learning architectures' pulling and improving generalization. The study of Zhang et al. (2024) confirms the suitability of using LANMAO neonatal sleep recorder as a low-channel alternative to full polysomnography, showing excellent agreement in EEG sleep scoring of neonates and high fidelity of the EEG signals. Consequently, the use of pediatric-friendly EEG acquisition systems is supported. This is a crucial factor for the development of future low-channel or wearable EEG solutions for detecting apnea and RERA.

Dritsas and Trigka (2024) were looking into a multi-class classification method for the automated prediction of sleep disorders and stated that the class imbalance issue and the selection of multiclass-appropriate metrics are very important. The results of their study give a clarifying guidance for the internal 3-class models that will be developed before they are mapped onto the binary challenge formats.

In agreement with this study, Hamouda et al. (2024) are proposing a multi-channel classification of sleep stages based on an ensemble learning framework that shows the relative robustness of combining different learners over a single one in noisy EEG conditions.

Deep-learning-based sleep staging is still getting developed, as with Jha et al. (2024) they introduced the SlumberNet a residual CNN model. Their model portrays how effective residual feature extractors in the capturing of time–frequency sleep EEG patterns while training (Jha et al., 2024) get released and also stabilizing the training process. In parallel, Famà et al. (2024) successfully employed EEG machine learning models to measure the conversion time and differentiate the subtypes in patients suffering from REM sleep behavior disorder. Although these two conditions are distinct when it comes to clinical criteria, their research demonstrates the combined use of temporal modeling and EEG-only biomarkers to aid prognostic processes, thereby indicating the vast clinical future of single-modality EEG pipelines.

Hidalgo Rogel et al. (2024) performed an investigation into the use of EEG for drowsiness detection in driving that utilized scalable machine learning models, pointing out lightweight design and inference efficiency as the main principles, which could be transferred directly to the monitoring of respiratory events in real-time or with wearable devices (Hidalgo Rogel et al., 2024). At the same time, El Hadiri et al. (2024) concentrated on studying non-linear EEG parameters like entropy and fractal measures within the framework of sleep-stage detection, and the conclusion was drawn that engineered nonlinear features can substantially improve the performance even more so in tasks with subtle EEG signatures.

A significant architectural change is presented by Guo et al. (2024), who have introduced the FlexSleepTransformer, a transformer-based staging system that can function with different EEG channel configurations. Their flexible-input design is especially applicable for reduced-channel or wearable apnea/RERA detectors.

Explainability, as well as uncertainty estimation, have become new priorities in the field. Heremans et al. (2024) by introducing U-PASS, an uncertainty-guided deep learning pipeline that quantifies the model's confidence for automated staging, have taken a step towards uncertainty-aware designs that are incorporated into clinical decision support systems and can be adapted to the workflows of apnea/arousal detection.

Generalization across datasets and scoring quality was addressed by Ganglberger et al. (2024), who paired transfer learning with scorability-weighted training to improve cross-cohort sleep-staging accuracy. Their findings reinforce the necessity of pretraining and label quality modeling, particularly the case in the transition from PSG to the wearable or pediatric cohorts.

On the other hand, Zaman et al. (2024) introduced SleepBoost; a tree-based ensemble with hierarchical feature extraction that is classical. The latter is still competitive while at the same time being of low computational cost, making it a good candidate for deployment, especially when used as a baseline against deep models (Zaman et al., 2024). Finally, Kolhar et al. (2025) made use of particle swarm optimization to optimize LSTM networks for CAP EEG and proved that evolutionary methods can still outperform deep learning techniques.

3. Methodology

To develop this research, the open source software Spyder IDE and the Google Colab platform were used, which allow scientific programming in Python in versions 3.6 and 3.7. Equipment and software used to develop the research:

a) SSH Power Shell Client graphical interface used to connect to the remote machine's GPU node and
b) Laptop used for remote connection and activities related to algorithm development

Given the complexity and computational cost of the implemented systems, two pieces of computing equipment were also used: the first was a Laptop Hp 440 Intel(R) Core(TM) for use. and the second was a remote machine owned by the Artificial Intelligence Laboratory (ICA/PUC Rio), which has a GeForce GTX 1080 Ti GPU with 8 GB of RAM, accessed using the SSH Secure Shell Client software.

3.1. Data base

The dataset used was The PhysioNet Computing in Cardiology Challenge 2018 (Hospitals et al., 2018), has the polysomnographic (PSG) records of 1,985 individuals. The database is partitioned into two groups: (i) 994 records for balanced training and (ii) 989 records for Test (whose classes are not in the public domain). Each of the polysomnographic recordings has a total of 13 physiological signals referring to the measurement of brain, muscular, cardiac and respiratory activity. These signals are listed in Table 1.

Table 1: Signals Contained in Each OF THE Polysomnographic Records in the Database

Channel	Signal [µV]	Definition
F3-M2	frontal activity	EEG
F4-M1	frontal activity	EEG
C3-M2	central activity	EEG
C4-M1	central activity	EEG
01-M2	later activity	EEG
02-M1	subsequent activity	EEG
E1-M2	left	EOG of the eye
Chin1-Chin2	µV	chin EMG

ABD	µV	EMG of abdominal movement
Chest	µV	EMG of Chest movement
Air flow	Unitless / arbitrary	Respiratory air flow measurement
SaO2	(%)	Oxygen Saturation
ECG	(mV)	Cardiac activity

In Figure 1, an example of a segment of the signals contained in one of the PSG records in the database is presented, in the red area are the EEG signals (i.e. channels 0-5, mentioned in Table 1), the yellow area corresponds to the EOG signal (i.e. Channel 6, mentioned in Table 1), the green area corresponds to the chin, chest and abdomen EMG signals (i.e. channels 7-9, mentioned in Table 1) and the celestial area corresponds to the air flow signals, SaO2 and ECG.

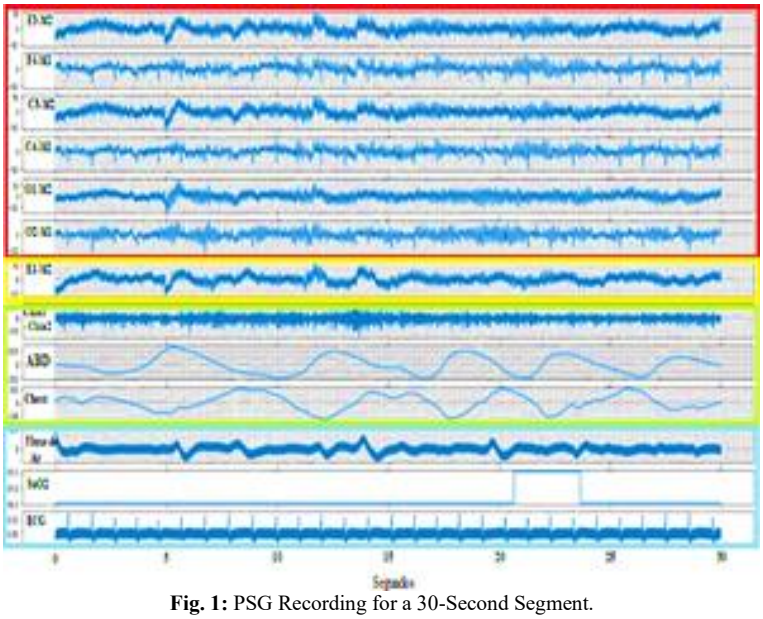


Fig. 1: PSG Recording for a 30-Second Segment.

The PSGs were divided into 30-second intervals, each interval has different types of observations (notes) made by MGH certified sleep technologists associated with the following categories: (i) sleep stages, (ii) Awakenings related to respiratory events and (iii) awakenings related to movement disorders such as bruxism and restless leg syndrome. In the case of sleep stages, six stages are considered: wakefulness, NREM1 stage, NREM2 stage, NREM3 stage, REM sleep and indefinite. As for the excitements, they were classified according to their type in the categories: spontaneous excitements, RERA, bruxisms, hypoventilations, hypo-apneas, apneas, vocalizations, snoring, periodic leg movements (PML), periodic Cheyne-Stokes breathing, or partial airway obstructions. Figure 2 shows the distribution of examples for each of the previously mentioned classes. It is noteworthy that the target observations for this research, corresponding to the segments with RERA events, events related to apnea/hypopnea and Normal are in a separate file from the previously mentioned observations.

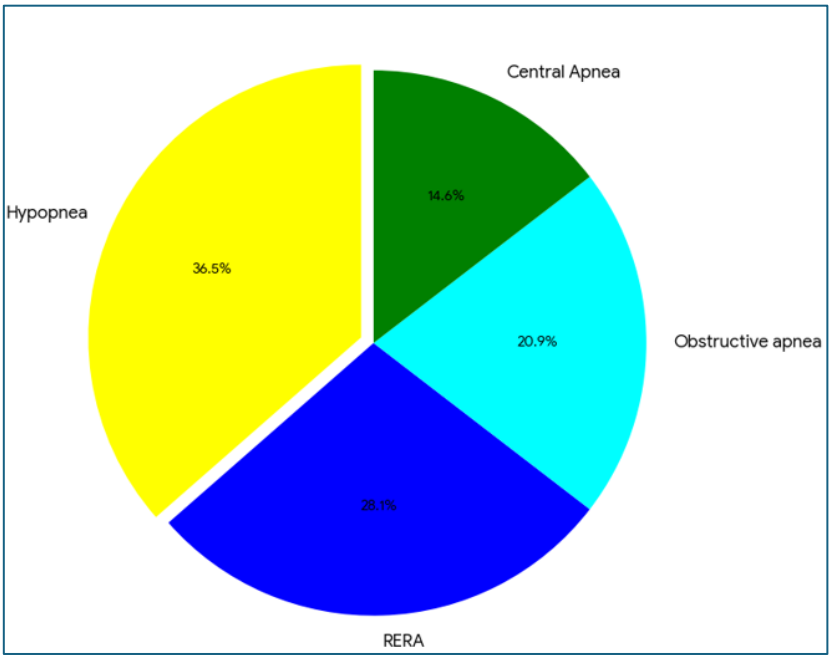


Fig. 2: Number of Examples for Each of the Sleep Disorders Present in the Database.

3.2. Methods

Figure 3 illustrates the methodology applied to develop this research, which is divided into three phases: (i) the study of the database, its organization, including an exploratory analysis and the understanding of the information contained in the records for its subsequent analysis; (ii) data preprocessing, feature extraction and evaluation of the proposed classification systems, and (iii) validation of results, as well as their subsequent analysis. Each of the phases is shown below.

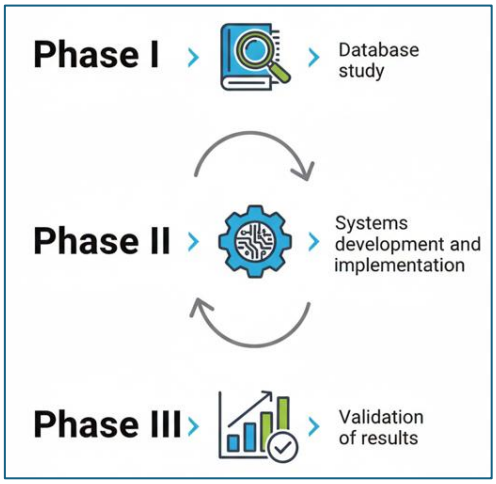


Fig. 3: Methodology Phases.

For this research, the data was divided into three subsets: (a) Training, (b) Test and (c) validation, with sizes equivalent to 70%, 20% and 10%, respectively. Then, electroencephalography (EEG) signals were extracted from each of the PSG records along with the class labels divided into: Normal, RERA Event and Events associated with apnea/hypopnea.

3.3. Normal

The class designates the absence of excitations as can be seen in Figure 4, in which an example of the normal class is presented for an epoch of 30 seconds of the 6 EEG channels, in which it is observed that none of the signals present abrupt changes throughout the epoch and each of the EEG signals presents normal wave patterns.

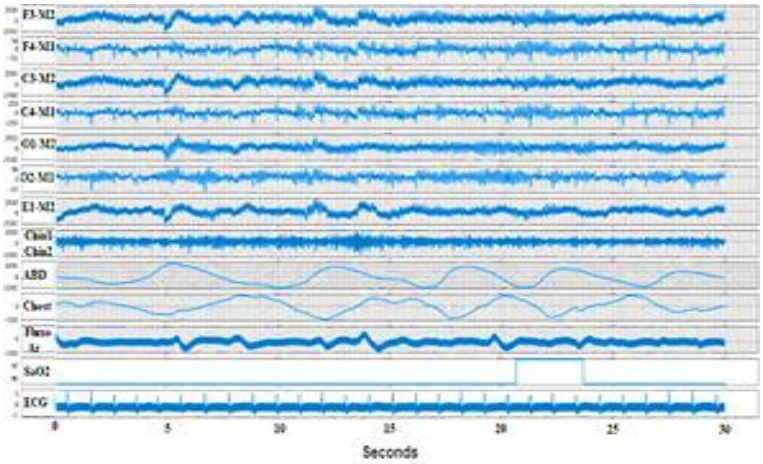


Fig. 4: Example of a Normal Class Polysomnography Segment.

3.4. RERA event

This class designates periods with awakenings caused by respiratory effort (Respiratory Effort-Related Arousal-RERA). In Figure 5 an example is presented for a 30-second epoch in which it is possible to observe some changes in the behavior of the signals between 15 and 30 seconds that are characteristic of RERA events.

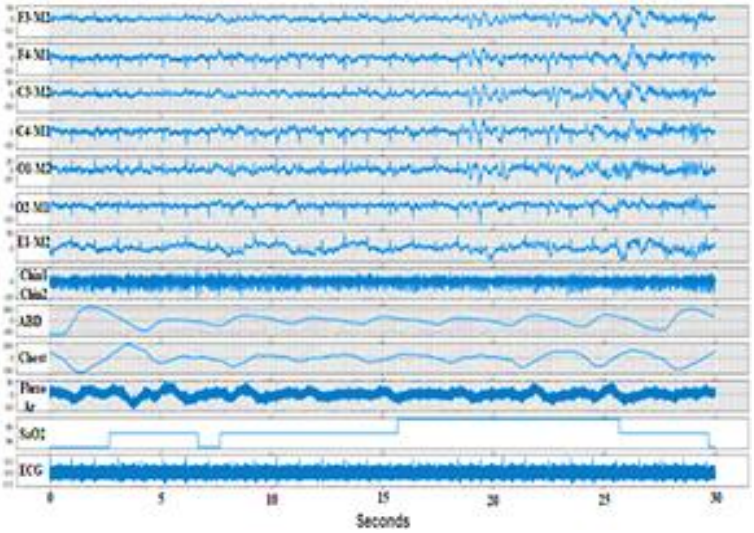


Fig. 5: Example of a Polysomnography Segment of a RERA Event.

3.5. Events associated with apnea/hypopnea

This class designates regions with respiratory events associated with apnea/hypopnea, in Figure 6 an example of this class is presented for a 30-second epoch of the 6 EEG channels, in which it is observed that for channels F3-M2, C3- M2, C4-M1 and 01-M2 between 10 and 20 seconds there are abrupt changes in the waveforms of these signals.

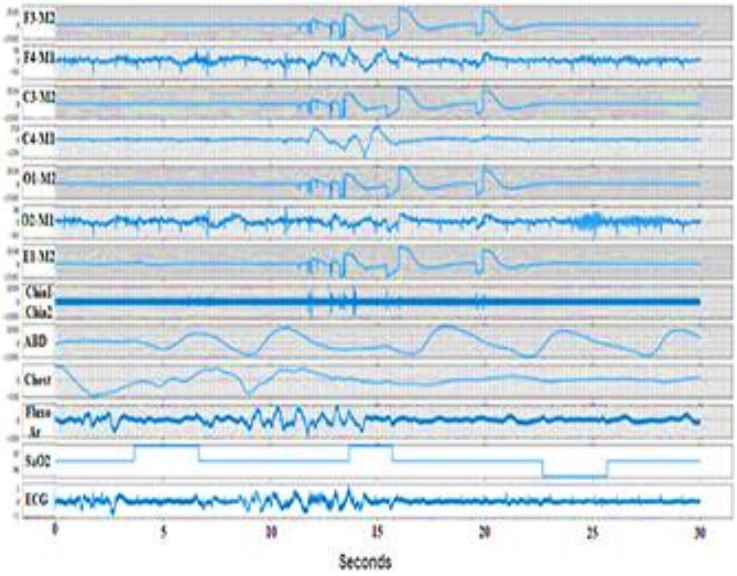


Fig. 6: Example of A Polysomnography Segment Associated with an Apnea/Hypopnea Event.

3.6. PHASE II: systems development and implementation

At this stage, five systems were presented based on the use of machine learning techniques and the extraction of features in the time and frequency domain: The first system is based on the use of methodologies conventionally used in EEG signal classification tasks and the remaining ones, in the use of CNN and LSTM networks and Class Weight and focal loss were evaluated to deal with unbalanced data.

3.7. Data notation

The EEG_raw for each individual are represented as a matrix of size [Ni, 6] and the class labels in a yi sequence of size Ni, where Ni is the number of samples in each individual's signals (length of signals). Because each system requires a specific input size, the dimensions of the records were modified according to the system. For system 1, the sequence of class labels yi was segmented into 30-second epochs, which results in a new sequence of class labels Yi with size ni, where ni is calculated from Equation 1.

$$n_i = \frac{N_i}{F_s * \omega} \tag{1}$$

Where,F_s is the sampling frequency and w is the window size in seconds (30 seconds). Then, for each of the six EEG signals present in EEG_raw, the average powers for the EEG frequency bands (Delta, alpha, Beta, Theta and Gamma) for each of the n_i epochs, thus resulting in a matrix of size [n_i, 30]. Table 2 presents a summary of the notations used for system 1.

Table 2: Data Notation for System 1

Name	Definition	Dimension
yi	Class labels sequence	Ni
Yi	Class labels by season	ni
EEG_raw	EEG time series	Ni x 6
X_EEG	PMB Matrix	ni x 30

In the case of systems 2, 3, 4 and 5, an axis was added to the input matrix (EEG_raw) generating a matrix of dimensions [1, Ni, 6]. After the new input matrix was segmented into epochs of 30 seconds, resulting in a signal matrix of size [1, ni, 6], however for, the yi class label sequence was encoded in One Hot Encoding format resulting in a matrix of size [yi, 3]. Finally, the data and class labels used are compressed into batches of [Bs, 1, (wFs), 6] for the EEG signals and [Bs, 3] for the class labels, where Bs is the size of the subset of input data to the networks at each step during training. Emphasizing that the number of steps refers to the total number of steps (batches of samples) before declaring that an epoch has ended and starting the next epoch, it is generally calculated from the Ni \ Bs relationship. Table 3 presents the summary of the notations used for these systems and Bs represents Batch size

Table 3: Data Notation for Systems 2, 3, 4 and 5

Name	Definition	Dimension
yi	Sequence class labels	Ni x 3
EEG_raw	EEG time series	1 x Ni x 6
EEG_Batch	Batch EEG time series	Bs,1,(wFs),6
Y_Batch	Batch class labels	Bs x 3

* One Hot Encoding Format.

3.8. System 1

In this system, a methodology conventionally used for feature extraction and EEG analysis is applied. Also in agreement with other works in the literature (Rodrigues et al., 2019; Kaya & Ertuğrul, 2018; Lv & Li, 2020; Cohen, 2014; Phan & Mikkelsen, 2022; Chiranjeevi & Dixit, 2017), three techniques widely used for classification tasks in EEG signals were chosen as classifier methods: (i) K-Nearest Neighbor (KNN), (ii) Naive Bayes (NB) and (iii) Random Forest (RF). Figure 7 shows the flow diagram implemented for the development of this system:

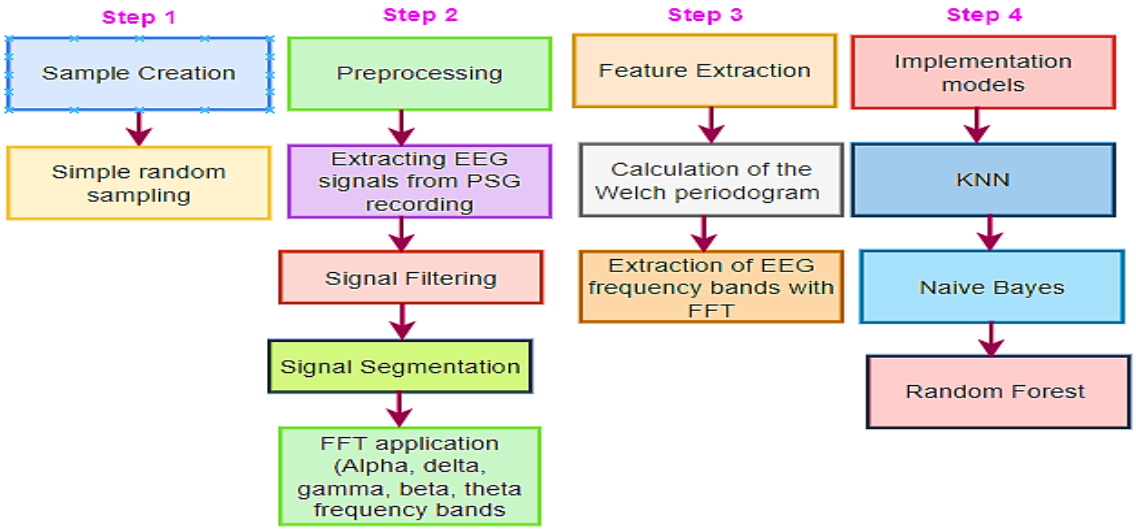


Fig. 7: Flow Diagram of the Steps for Developing and Implementing System 1.

- Sample creation: a representative random sample of the data set of size n was created, where n was calculated using Equation 2 due to computational limitations for processing the complete data set.

$$n = \frac{N_{pq}K^2}{e^2(N-1)+pqK^2} \quad (2)$$

In the previous expression, n is the sample size, N is the population size, K is the critical Z value or desired confidence level, e is the level of absolute precision, p is the approximate proportion of the phenomenon under study in the reference population and q is the proportion of the reference population that does not have the study phenomenon (1 − p). To ensure representativeness, a confidence interval of 95% was chosen, a margin of error of 5%, the population size was 994 individuals and p and q had values of 0.5. The result of this calculation is presented in Equation 2, thus resulting in a calculated sample size of 278 records.

- Pre-processing: From each polysomnographic recording (PSG) the signals corresponding to the electroencephalographic derivations were extracted: F3-M2, F4-M1, C3-M2, C4-M1, 01-M2 and O2-M1 (channels 0 to 5 of Table 1), represented in Figure 8. Then, each signal was filtered using a bandpass FIR filter in the frequency range of 0.1 to 45 Hz and segmented into 30-second epochs, following AASM recommendations. Then each epoch was transformed to the frequency domain using Fast Fourier Transform (FFT).

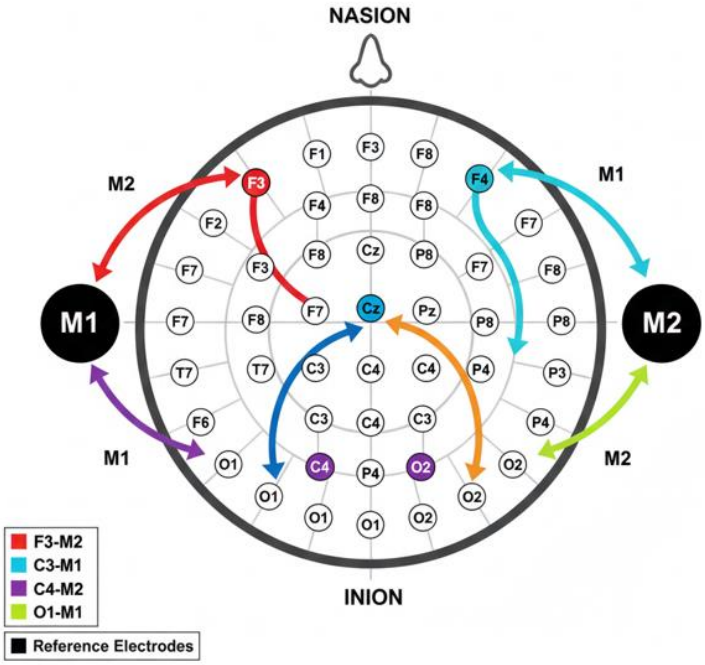


Fig. 8: EEG Derivations Contained in the PSG Records of the Database.

- Feature Extraction: For this step, the power spectral density (PSD) was first calculated using the Welch periodogram method. To calculate the periodogram, a window of 4 seconds was used as the use of this method allows to drastically reduce the temporal variation of the EEG signal compared to the classic periodogram.

After calculating the Welch periodogram for each segment (Figure 9a), the frequencies of interest corresponding to the delta, theta, alpha, beta and gamma bands are extracted. The average power of each EEG band is then calculated. Average band power (average bandpower), is a relevant metric in the analysis of EEG signals, it consists of calculating the summary of the contribution of each frequency band (delta, alpha, beta, theta and gamma), that is, the area under the curve of each band of the periodogram. An example is the blue region in Figure 9b. To approximate this area, Simpson's composite rule is used.

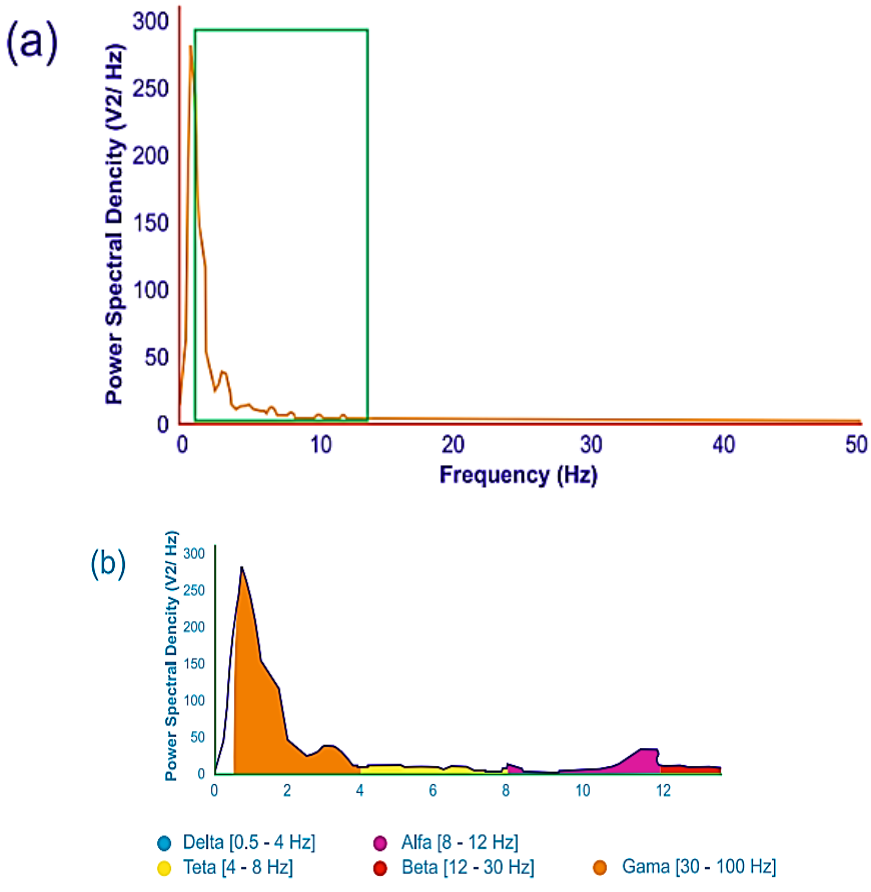


Fig. 9: Welch Periodogram: (A) Graph of the Welch Periodogram, the Power Spectral Density (PSD) Usually Measured in µv 2/Hz for EEG Data. (B) Welch Periodogram Divided into Its Delta, Theta, Alpha, Beta and Gamma Frequency Bands.

Finally, the average powers of each EEG frequency band and the corresponding classes were organized in a matrix of size [ni, 31], with ni being the number of 30-second epochs in the polysomnographic record, as illustrated in Figure 10.

	F3M2_Delta	F3M2_Theta	F3M2_Alfa	F3M2_Beta	F3M2_Gamma	F3M2_Delta	.....	Labels
0	101.8518519	234.3915344	216.6666667	41.53439153	48.14814815	141.5343915	.....	0
1	154.7619048	169.047619	238.3597884	210.3174603	157.6719577	210.8465608	.....	0
2	57.93650794	129.6296296	103.1746032	201.0582011	163.7566138	107.6719577	.....	1
3	70.63492063	32.8042328	25.92592593	107.6719577	116.1375661	200.2545503		
4	88.35978836	168.2539683	201.8518519	133.5978836	217.7248677	107.4074074		
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
ni	94.44444444	154.021164	164.5502646	79.36507937	238.6243386	157.6719577	.....	1

Fig. 10: Polysomnographic Data Matrix for an Individual.

- Model implementation: Once the previous steps have been completed, the matrix with the average band powers and the class labels are used to train the classification models: (i) KNN, (ii) NBand (iii) Random Forest.

3.9. System 2: CNN1

This system is composed of a network architecture, formed by three convolutional blocks (convolutional and Max-pooling sampling layers), as illustrated in Figure 11a.

3.10. System 3: RCNN1

Figure 11b illustrates the network architecture of system 3, in which, first, a CNN network processes the 30-second epochs of EEG data (non-overlapping data windows), generating a feature map that represents input sequence.

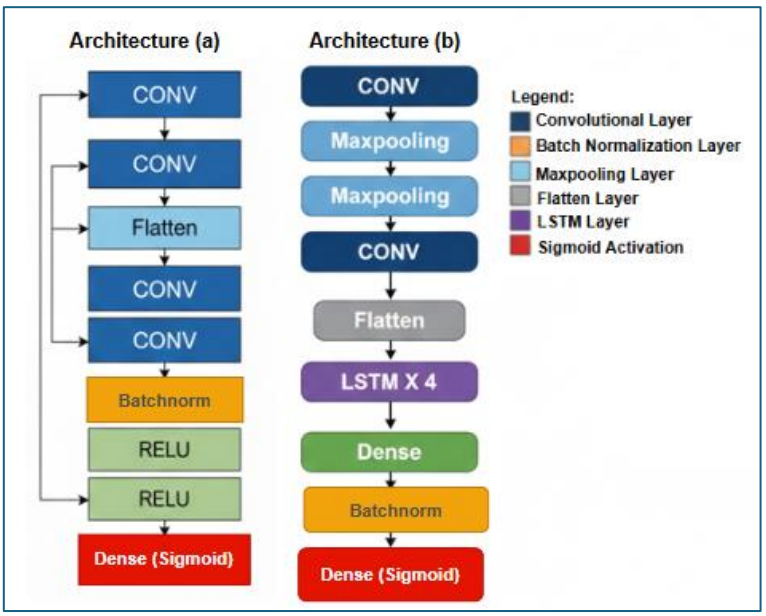


Fig. 11: Architectures used for (a) System 2 and (b) System 3.

3.11. System 4: CNN2

This system is a variation of system 2, in which the addition of CBR convolutional blocks is proposed.

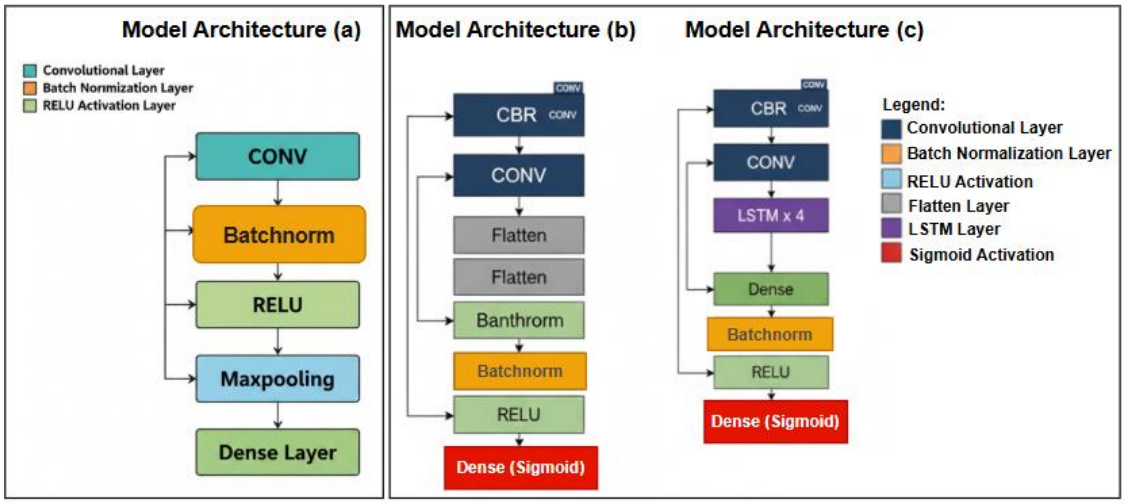


Fig. 12: (A) CBR Block and Architectures Used for (B) System 4 and (C) System 5.

The architecture used in this system is composed of two CBR blocks, followed by a convolutional layer (CL), a fully connected layer, a BatchNormalization layer, a ReLu function and a sigmoid function to generate the prediction of respiratory events, as illustrated in Figure 12b.

3.12. System 5: RCNN2

This system is a variation of system 4, in which, followed by the CBR blocks and the CL, four stacked LSTM networks are added in charge of maintaining the temporal relationship of the spatial features extracted in the CBR Blocks and the immediately previous CL. The architecture used in this system is illustrated in Figure 12c.

3.13. PHASE III: validation of results

In this phase, different tests were carried out with all systems, aiming to find the best combination of parameters in each of them. To analyze the ability to discriminate and generalize in the face of new data, the validation and test sets were evaluated and the metrics Confusion matrix with True Positives (TP), True Negatives (TN), False Positives (FP), and Falses Negative (FN), Area Under the Receiver Operating Characteristic Curve (AUROC), Precision, Recall, Area under the Precision Recall curve (AUPRC), Accuracy, Specificity were calculated.

4. Results and Discussion

This section presents and describes the results obtained in the development and implementation phases of the proposed systems and the validation of results, for the identification of awakenings related to RERA events and apnea/hypopnea in PSG recordings based only on EEG data.

4.1. System 1 results

Table 4 presents the results obtained using the RFalgorithm as a classifier. Table 7.2 presents the results obtained using the NBalgorithm as a classifier. In Table 7.3, the results obtained using the KNN algorithm as a classifier are presented. The best results were with the use of NBtechniques (AUROC: 0.5241 and AUPRC: 0.1254) and KNN (AUROC: 0.5019 and AU-PRC: 0.4494). Given the complexity in detecting RERA events, the results are relatively low, however, studies such as (Shoeb & Sridhar (2018) also presented low results (close to those obtained in this work), using the same database (Physionet Challenge 2018). Considering the results obtained in this system 1, the possibility of employing more robust and efficient techniques based on DL (systems 2, 3, 4 and 5) was evaluated.

Table 4: Results Obtained by System 1 Using Three Techniques.

Random Forest			Naive Bayes			KNN		
# Estimators	AUROC	AUPRC	# Estimators	AUROC	AUPRC	# Neighbors	AUROC	AUPRC
10	0.3769	0.0081	1−10	0.5072	0.1115	1	0.4164	0.0859
60	0.6125	0.0789	1−9	0.4796	0.1645	2	0.5019	0.2519
70	0.36125	0.0789	1−8	0.5124	0.1124	3	0.4341	0.4351
70	0.36354	0.0812	1−7	0.4423	0.1264	4	0.4365	0.4394
80	0.34897	0.1134	1−6	0.5019	0.1164	5	0.4461	0.4419
90	0.36354	0.0812	1−5	0.4721	0.1165	6	0.4456	0.4495
100	0.45126	0.1892	1−4	1.4842	0.1198	7	0.4461	0.4449
100	0.36354	0.0821	1−3	0.5241	0.1199	8	0.4461	0.4449
100	0.36254	0.0821	1−2	0.5124	0.1254	9	0.4615	0.4449
100	0.36425	0.0821	1−1	0.5426	0.1159	10	0.4987	0.4494
100	0.34798	0.1082	1	0.4246	0.1153	11	0.4095	0.4051
200	0.36554	0.0812				[1-50]	0.4094	0.4055

4.2. System 2 results

For this system, tests were carried out with different numbers of CLs, number of neurons and kernel sizes. The best combinations of these hyperparameters are presented in Table 5. For this network architecture, 2D CLs (conv2d) were used, which allowed the signals to be operated in two dimensions to extract the greatest number of features possible. For each of the sampling layers (Max-pooling), kernels of size (1, 2) were used, Adam was used as the optimizer and the loss functions were alternated between focal loss and binary crossentropy. The number of neurons in the FC layer is in column #D, and the batch sizes used are in the BS column. Finally, the DBT column expresses the data balancing technique used: i)class weight (cw) and ii) focal loss (fl). For each of the network configurations presented in Table 5, the metrics were calculated.

Table 5: Hyperparameter Settings for System 2

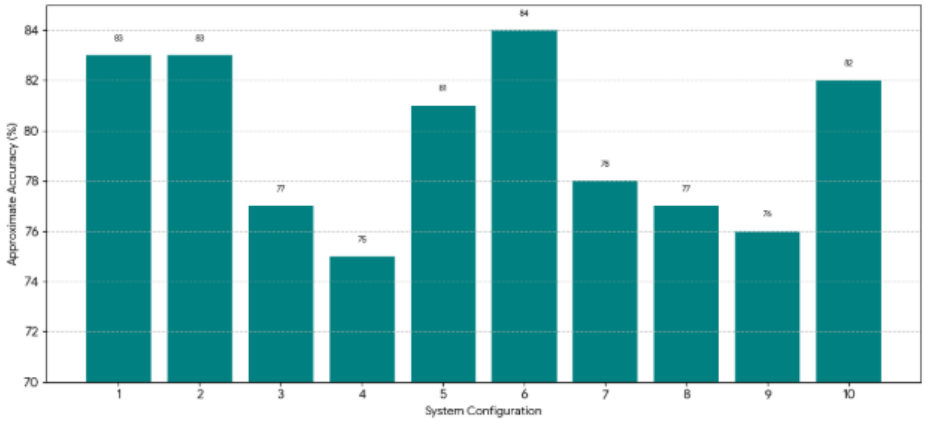
#	CNN*	Neurons	Kernels	# D	BS	DBT
1	4	32,64,128,256	3,3,3,3	128	128	–
2				128	256	–
3				128	128	fl
4		16,32,64,128		64	128	fl
5		32,64,128,256		128	128	cw
6		16,32,64,128		64	28	w
7		16,32,64	7,5,3	64	128	fl
8		32,64,128,256	11,7,5,3	128	128	fl
9		16,32,64,128		64	128	cw
10		32,64,128,256	7,5,3,3	28	28	fl

Table 6: Results Obtained Evaluating System 2.

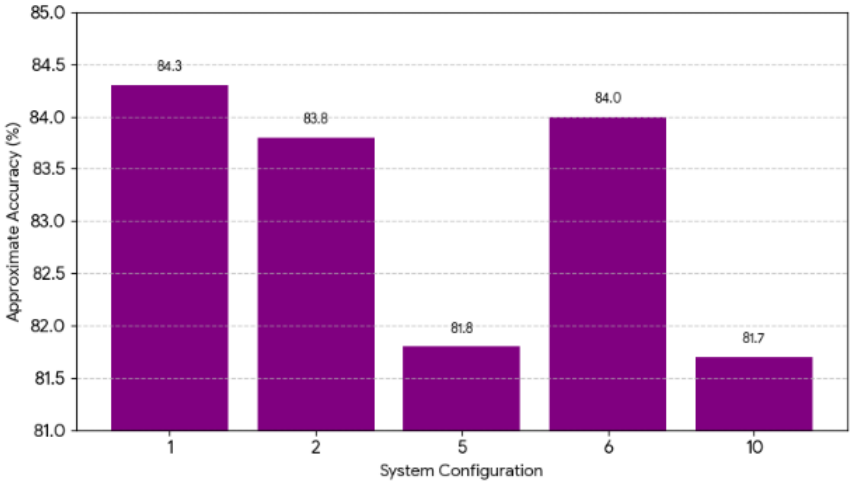
#	Accuracy Train.	Val	Test	AUROC*	AUPRC*	Normal*	RERA* (%)	ApHip* (%)	WCE**
1	0.8890	0.6926	0.7783	0.8293	0.7518	94.7060	64.7465	59.4364	1544735
2	0.8622	0.7308	0.7731	0.8202	0.7401	93.1167	26.2961	71.8766	1535848
3	0.7363	0.6841	0.7037	0.8176	0.7191	90.8964	49.2063	44.3569	1407835
4	0.7057	0.6812	0.6883	0.7971	0.6640	94.6466	21.5762	35.6562	1379465
5	0.8312	0.7153	0.7530	0.8029	0.7257	98.7869	32.4860	47.2165	1498245
6	0.7996	0.7501	0.7757	0.8238	0.7494	92.9562	69.1767	61.9261	1540065
7	0.7515	0.6499	0.7177	0.8208	0.7156	86.8365	42.7364	62.3963	1433605
8	0.7195	0.6385	0.7012	0.8056	0.6869	86.3863	15.3366	64.3862	1403195
9	0.8074	0.7539	0.6911	0.8284	0.7552	86.2862	17.5369	60.0760	1384655
10	0.7619	0.6614	0.7523	0.8279	0.7215	87.1364	45.7062	74.4267	1497035

* Calculated for the Test set; ** ApHip: apnea/hypopnea; TEC: Total Examples per Class; WCE: Total Well-Classified Examples.

Table 6 presents the results obtained for the Accuracy calculation for all Training, Validation and Test subsets, the AUROC and AUPRC values for the Test set, and the number of well-classified examples for each of the classes (Normal, RERA event and apnea/hypopnea). In Figure 13a, the comparison of the Accuracy values obtained for each configuration of system 2 evaluating the test set is presented. Once the results of the configurations were close, an enlargement was carried out in the range 80 to 85%, shown in Figure 13b. In this figure it is possible to clearly differentiate the best results (i.e. configurations 1, 2, 5, 6 and 10) in terms of Accuracy.



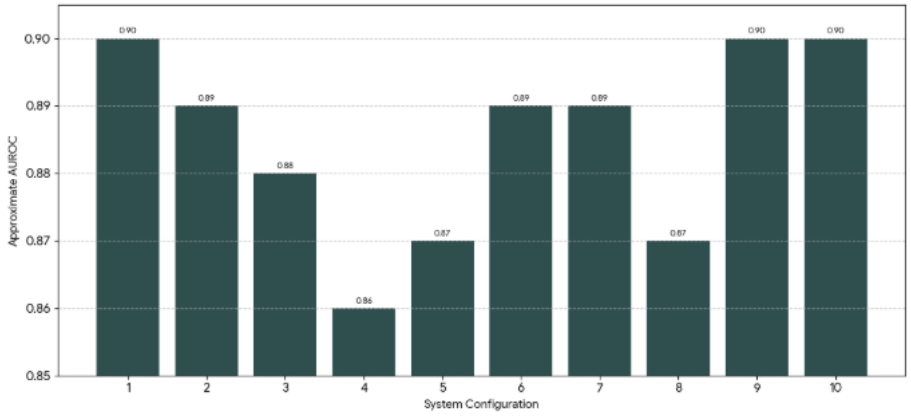
(A)



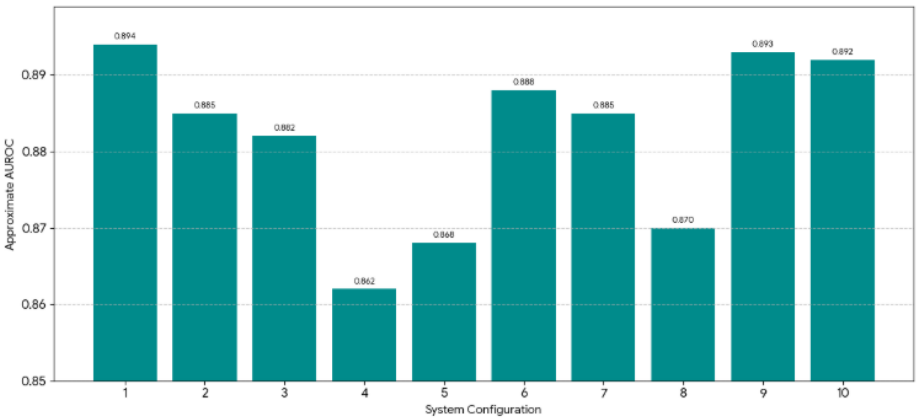
(B)

Fig. 13: Accuracy Values Obtained for System 2: (A) Comparison of the Accuracy Achieved by Each Hyperparameter Configuration and (B) Zoom of the Best Accuracy Results Obtained for the System.

In Figure 14a, the comparison of AUROC values obtained for each configuration of system 2 evaluating the test set is presented. As the configuration results were close, an enlargement was carried out in the range 0.85 to 0.90, shown in Figure 14b. In this figure it is possible to clearly differentiate the best results (i.e. configurations 1, 2, 6, 9 and 10) in terms of AUROC. In Figure 15a, the comparison of AUPRC values obtained for each configuration of system 2 evaluating the test set is presented. Since the results of the configurations were close, an expansion was carried out in the range 0.75 to 0.85, shown in Figure 15b. In this figure it is possible to clearly differentiate the best results (i.e. configurations 1, 2, 6 and 9) in terms of AUPRC.

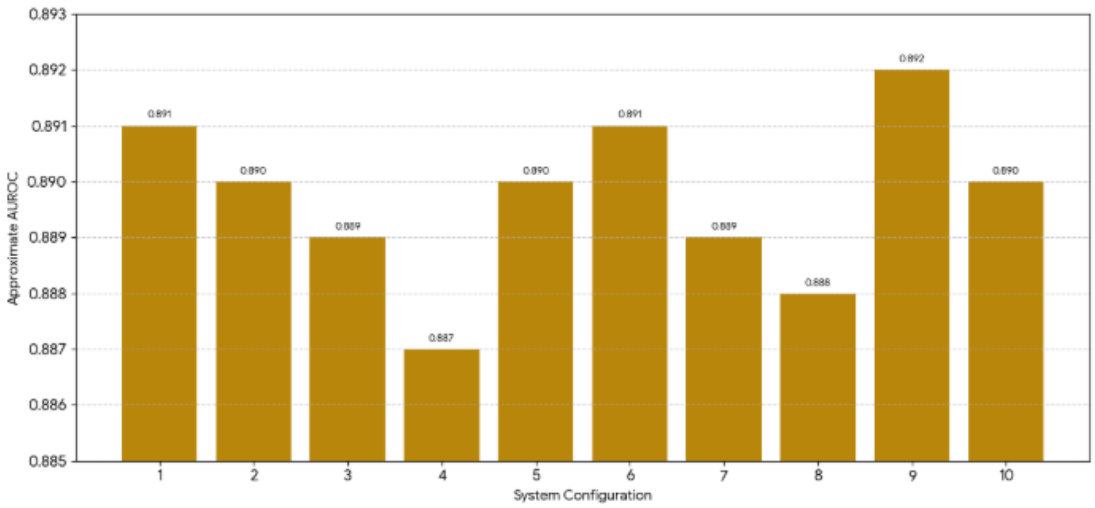


(A)

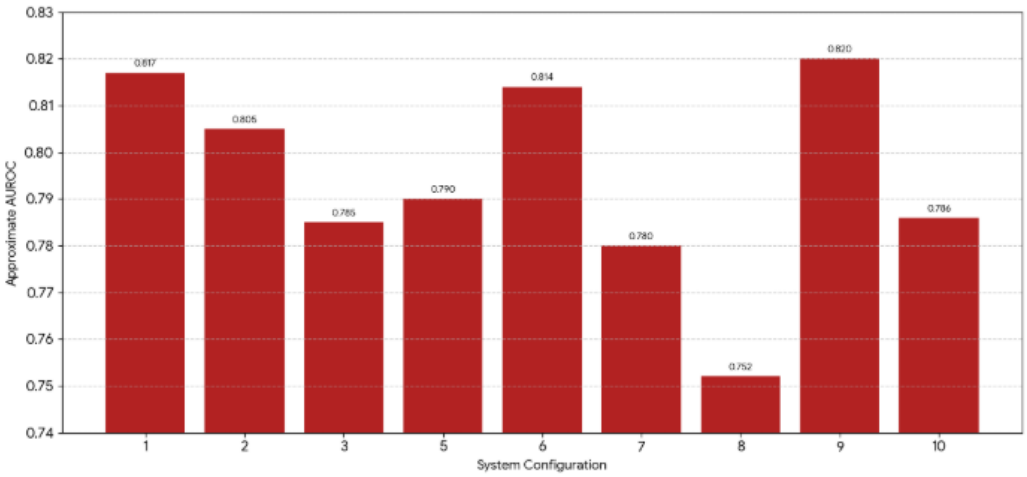


(B)

Fig. 14: AUROC Values Obtained for System 2: (A) Comparison of the AUROC Achieved by Each Hyperparameter Configuration and (B) Zoom of the Best AUROC Results Obtained for the System.



(A)



(B)

Fig. 15: AUPRC Values Obtained for System 2: (A) Comparison of the AUPRC Achieved by Each Hyperparameter Configuration and (B) Zoom of the Best AUPRC Results Obtained for the System.

4.3. System 3 results

For this system, tests were carried out with different numbers of CLs, kernel sizes and number of neurons. The best combinations of these hyperparameters are presented in Table 7. 1D CLs (conv1d) were used, which are used for extracting features in time series or sequential data, for each of the sampling layers (Max-pooling), kernels of size (2) were used as an optimizer Adam was used and the loss functions were alternated between focal loss and binary crossentropy. The number of neurons in the FC layer and in the LSTM networks is listed in column # D and U respectively. The batch sizes used in the BS column, finally, the DBT column expresses the data balancing technique used: i) class weight (cw) and ii) focal loss (fl).

Table 7: Hyperparameter Settings for System 3

#	CNN*	Neurons	kernels	LSTM	U	# D	BS	TBD
1		32,64,128	3,3,3	4	100, 100, 100, 100	64	256	-
2							128	Fl
3			7,5,3				256	
4	3							
5			3,3,3	3			128	Cw
6				2	128, 128	128	256	
7			7,5,3				128	
8			11,7,5				256	
9	9	16,32,64	3,3,3,3	4	100, 100, 100, 100	64	128	–

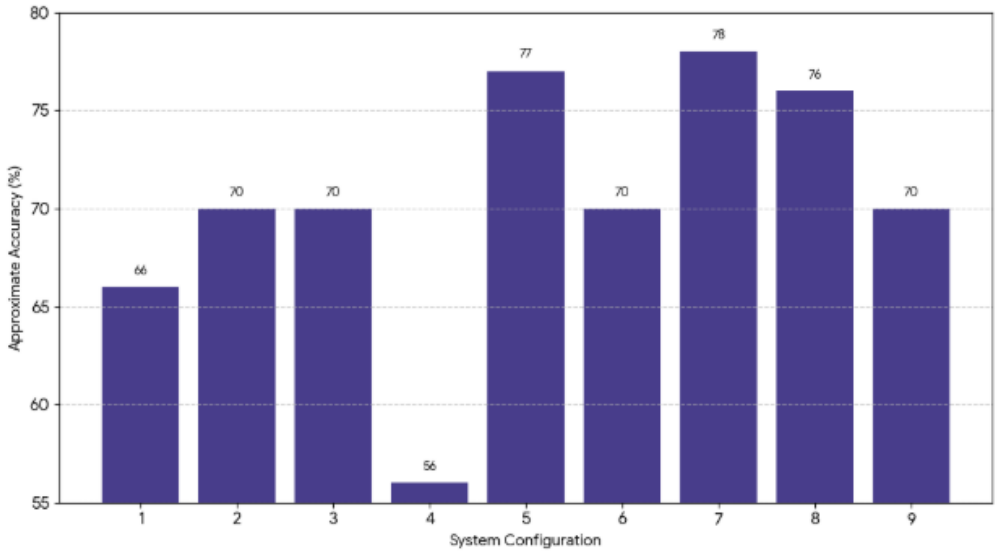
Table 8: Results Obtained Evaluating System 3

#	Accuracy Train.	Val	Test	AUROC*	AUPRC*	Normal*	RERA* (%)	ApHip* (%)	WCE**
1	0.6626	0.1536	0.6006	0.6821	0.6626	71.2861	18.5867	63.8663	107070
2	0.6377	0.3462	0.6482	0.7151	0.6377	80.6166	34.9866	53.6366	115910
3	0.6867	0.6367	0.6472	0.5159	0.6867	71.8468	31.8566	77.4363	115736
4	0.6846	0.2311	0.5055	0.5492	0.6846	65.8669	30.3362	37.9467	89720
5	0.7766	0.5832	0.7130	0.8043	0.7766	94.1567	15.9063	48.2161	127816
6	0.7676	0.7012	0.6487	0.7846	0.7673	72.8265	32.6868	75.1962	116006
7	0.7846	0.5502	0.7253	0.8123	0.7843	83.8661	70.6469	65.8468	130058

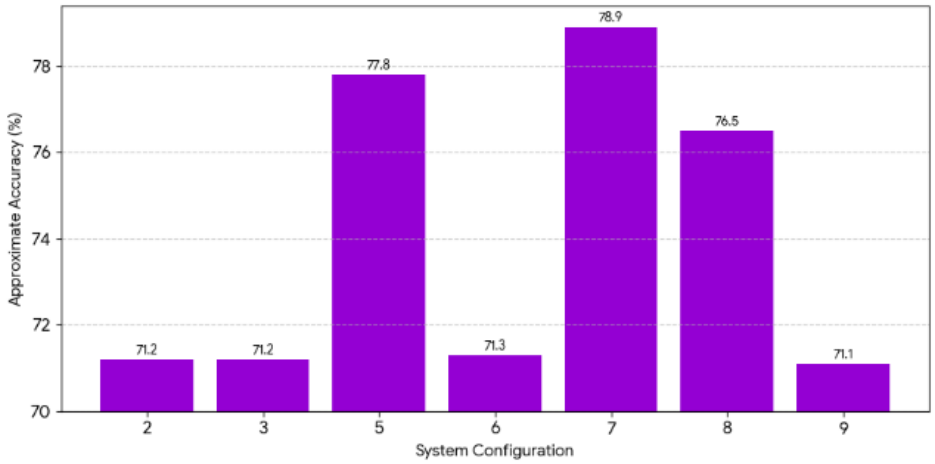
8	0.7825	0.7258	0.7004	0.8037	0.7825	98.3464	56.9061	21.1869	125496
9	0.7258	0.6634	0.6469	0.6921	0.7258	91.4262	24.6764	27.1063	115678
					TEC**	100	100	100	168611
1	0.6626	0.1536	0.6006	0.6821	0.6626	71.2861	18.5867	63.8663	107070

For each of the network configurations presented in Table 7, the metrics mentioned in section 6.2.3 were calculated. Table 8 presents the results obtained for calculating the Accuracy of all subsets of Training, validation and Testing; also, the AUROC and AUPRC values for the Test set; and the number of well-classified examples for the Normal, RERA Event and apnea/hypopnea classes. Finally, the WCE column shows the total number of examples that were well classified.

In Figure 16a, the comparison of the Accuracy values obtained for each configuration of system 3 evaluating the test set is presented. Once the results of the configurations were close, an enlargement was carried out in the range 70 to 80%, shown in Figure 16b. In this figure it is possible to clearly differentiate the best results (i.e. configurations 5, 7 and 8) in terms of Accuracy.



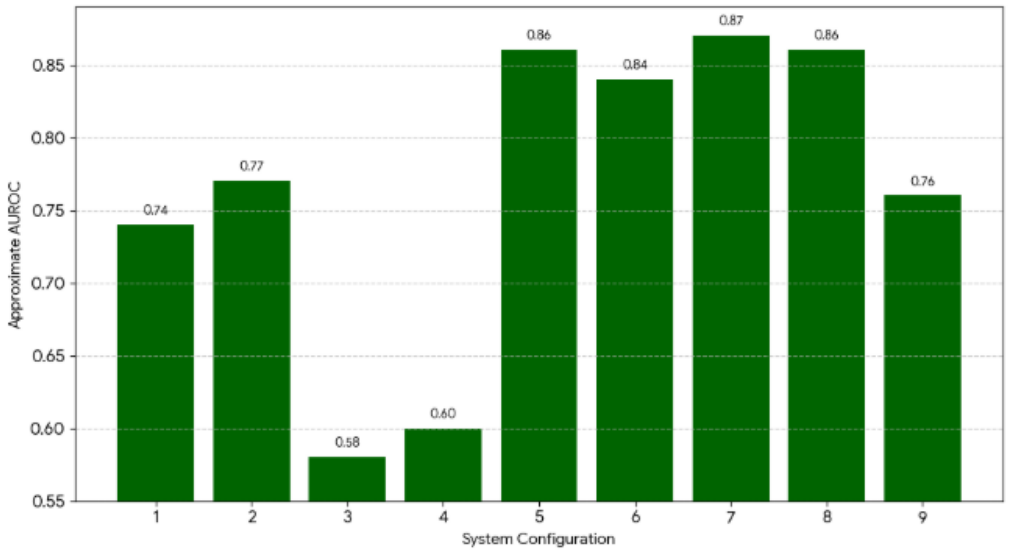
(A)



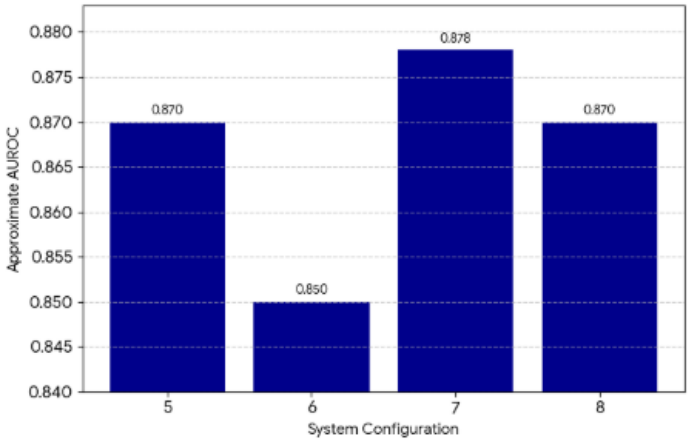
(B)

Fig. 16: Accuracy Values Obtained for System 3: (A) Comparison of the Accuracy Achieved by Each Hyperparameter Configuration and (B) Zoom of the Best Accuracy Results Obtained for the System.

In Figure 17a, the comparison of AUROC values obtained for each configuration of system 3 evaluating the test set is presented. As the results of the configurations were close, an expansion was carried out in the range 0.80 to 0.90, shown in Figure 17b. In this figure it is possible to clearly differentiate the best results (i.e. configurations 5, 7 and 8) in terms of AUROC. In Figure 18a, the comparison of AUPRC values obtained for each configuration of system 3 evaluating the test set is presented. Due to the configuration results being close, an enlargement was carried out in the range 0.70 to 0.80, shown in Figure 18b. In this figure it is possible to clearly differentiate the best results (i.e. configurations 5, 7 and 8) in terms of AUPRC.

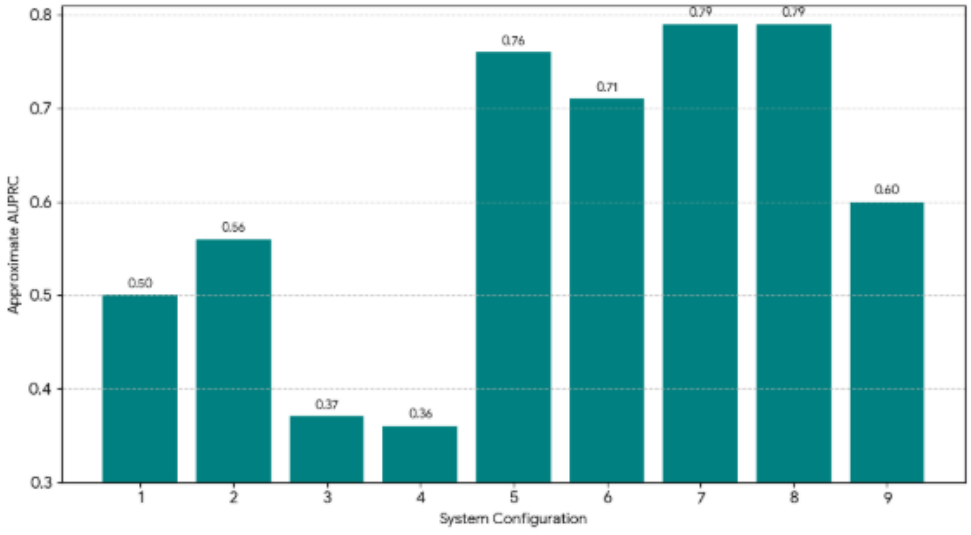


(A)

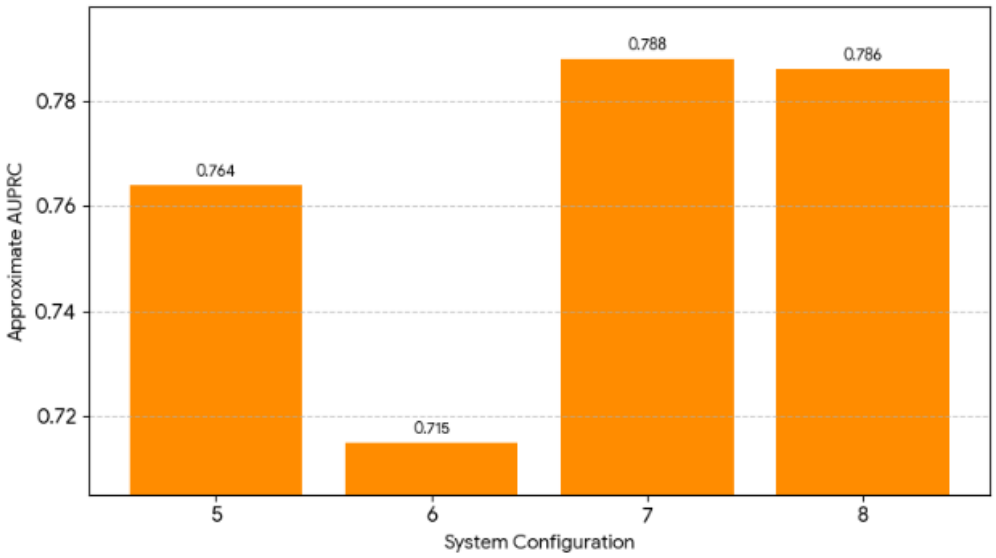


(B)

Fig. 17: AUROC Values Obtained for System 3: (A) Comparison of the AUROC Achieved by Each Hyperparameter Configuration and (B) Zoom of the Best AUROC Results Obtained for the System.



(A)



(B)

Fig. 18: AUPRC Values Obtained for System 3: (A) Comparison of the AUPRC Achieved by Each Hyperparameter Configuration and (B) Zoom of the Best AUPRC Results Obtained for the System.

4.4. System 4 results

For this system, tests were carried out with different numbers of CLs, number of neurons and kernel sizes. The best combinations of these hyperparameters are presented in Table 9. For this network architecture, 2D CLs (conv2d) were used, which allowed the signals to be operated in two dimensions to extract the greatest number of features possible. For each of the sampling layers (Max-pooling), kernels of size (1, 2) were used, Adam was used as the optimizer and the loss functions were alternated between focal loss and binary crossentropy. The number of neurons in the FC layer is shown in column #
D. The batch sizes used in the BS column, finally, the DBT column expresses the data balancing technique used: i) class weight (cw) and ii) focal loss (fl). For each of the configurations network presented in Table 9, the metrics were calculated.

Table 9: Hyperparameter Settings for System 4

#	CNN*	Neurons	Kernels	# D	BS	DBT
1		32,64,128,256	3,3,3,3	128	128	–
2						fl
3						wc
4			11,7,5,3			–
5	4	16,32,64,128	3,3,3,3			–
6						fl
7				64		cw
8			7,5,3,3	128	256	
9			11,7,5,3			
10					128	

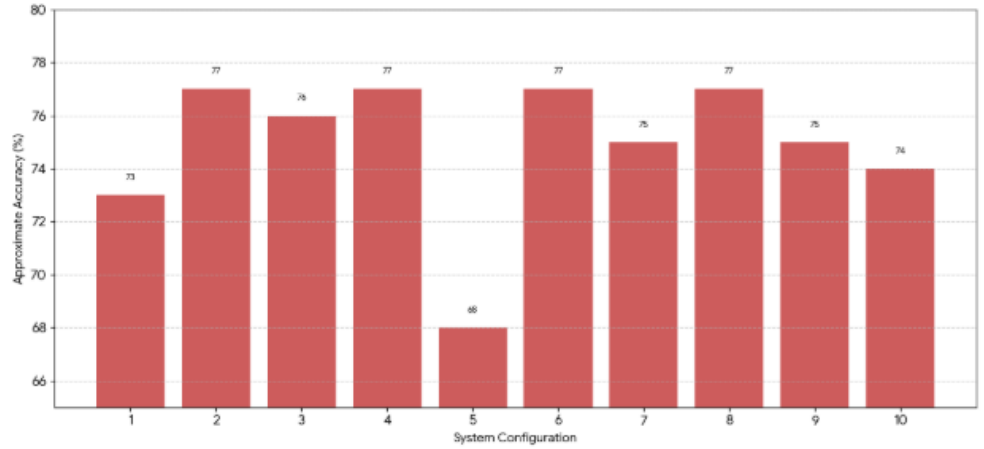
∗ CNN: Number of CNN layers; # D: Dense layer neurons; BS:Batch Size; DBT: Data Balancing Technique.

Table 10: Results Obtained Evaluating System 4

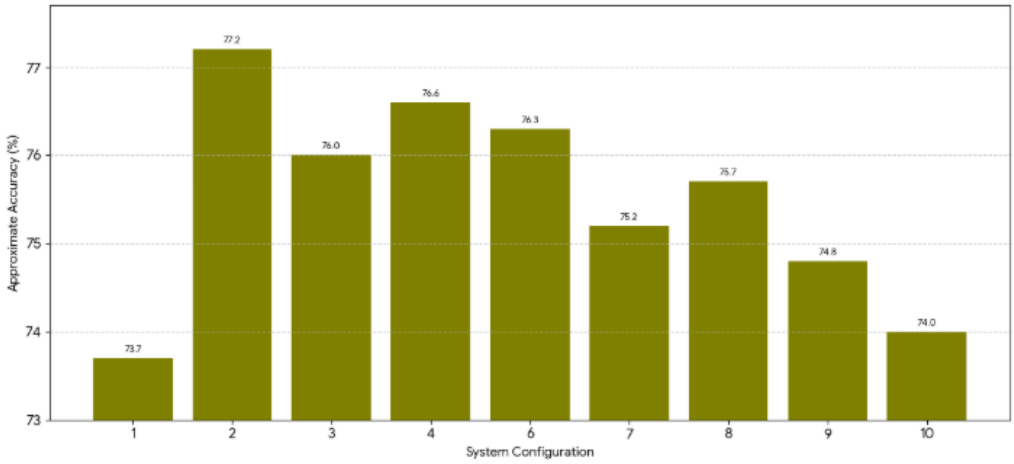
#	Accuracy Train.	Val	Test	AUROC*	AUPRC*	Normal*	RERA* (%)	ApHip* (%)	WCE**
1	0.8150	0.7254	0.6718	0.8132	0.7296	82.8063	11.8826	63.2263	133245
2	0.7356	0.6742	0.7068	0.8205	7861.93	93.1164	1.3528	52.3762	139669
3	0.8588	0.7102	0.6949	0.8231	0.7432	89.9261	8.7325	54.2066	137485
4	0.7109	0.6815	0.7009	0.8214	0.7376	93.0762	14.7322	46.5860	138586
5	0.7256	0.6534	0.6222	0.7738	0.6950	97.5965	10.9229	4.5761	124141
6	0.7197	0.7013	0.6981	0.8085	0.6770	91.7466	11.2023	49.9752	138072
7	0.8364	0.7364	0.6875	0.8019	0.7059	85.2268	11.7424	62.9965	136127
8	0.8955	0.7302	0.6920	0.8218	0.7350	84.7765	7.0921	62.9967	136953
9	0.5640	0.5760	0.6825	0.7899	0.6980	86.2862	4.2422	60.2068	135209
10	0.5914	0.5893	0.6748	0.8016	0.7196	85.5269	30.2325	52.2762	133796
					TEC**	100	100	100	181611

* Calculated for the Test set; ** Ap/Hyp: apnea/hypopnea;TEC: Total Examples per Class; WCE: Total Well-Classified Examples.

Table 10 presents the results obtained for calculating the Accuracy of all Training, Validation and Test subsets; also, the AUROC and AUPRC values for the Test set, and the number of well-classified examples for the Normal, RERA Event and apnea/hypopnea classes. Finally, the WCE column shows the total number of examples that were well classified. In Figure 19a, the comparison of the Accuracy values obtained for each configuration of system 4 evaluating the test set is presented. Once the results of the configurations were close, an enlargement was carried out in the range 70 to 80%, shown in Figure 19b. In this figure it is possible to clearly differentiate the best results (i.e. configurations 2, 4 and 6) in terms of Accuracy.



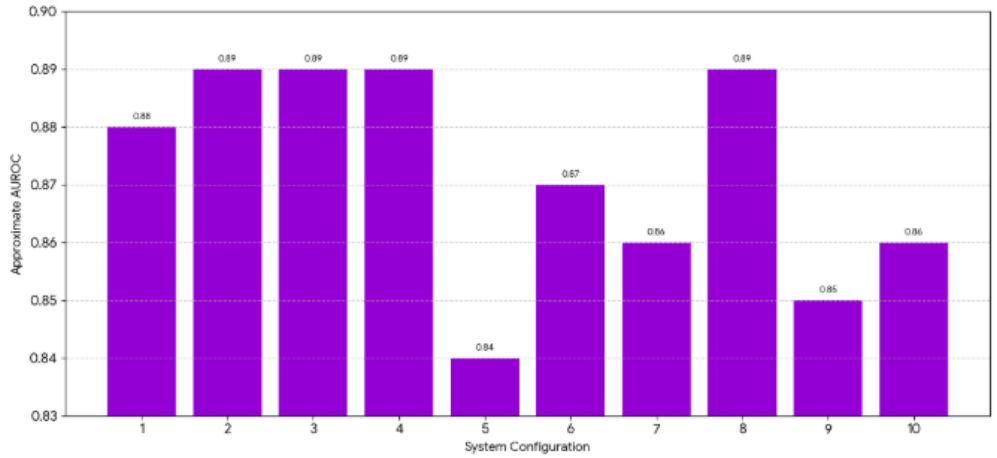
(A)



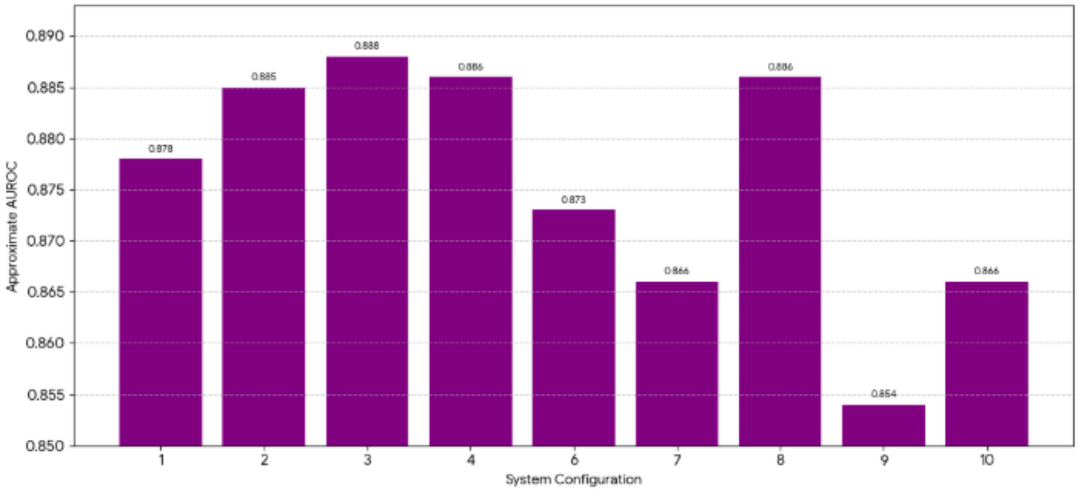
(B)

Fig. 19: Accuracy Values Obtained for System 4: (A) Comparison of the Accuracy Achieved by Each Hyperparameter Configuration and (B) Zoom of the Best Accuracy Results Obtained for the System.

In Figure 20a, the comparison of AUROC values obtained for each configuration of system 4 evaluating the test set is presented. As the configuration results were close, an enlargement was carried out in the range 0.85 to 0.95, shown in Figure 20b. In this figure it is possible to clearly differentiate the best results (i.e. configurations 2, 3, 4 and 8) in terms of AUROC. In Figure 21a, the comparison of AUPRC values obtained for each configuration of system 4 evaluating the test set is presented. Since the results of the configurations were close, an expansion was carried out in the range 0.72 to 0.82, shown in Figure 21b. In this figure it is possible to clearly differentiate the best results (i.e. configurations 1, 3, 4 and 8) in terms of AUPRC.

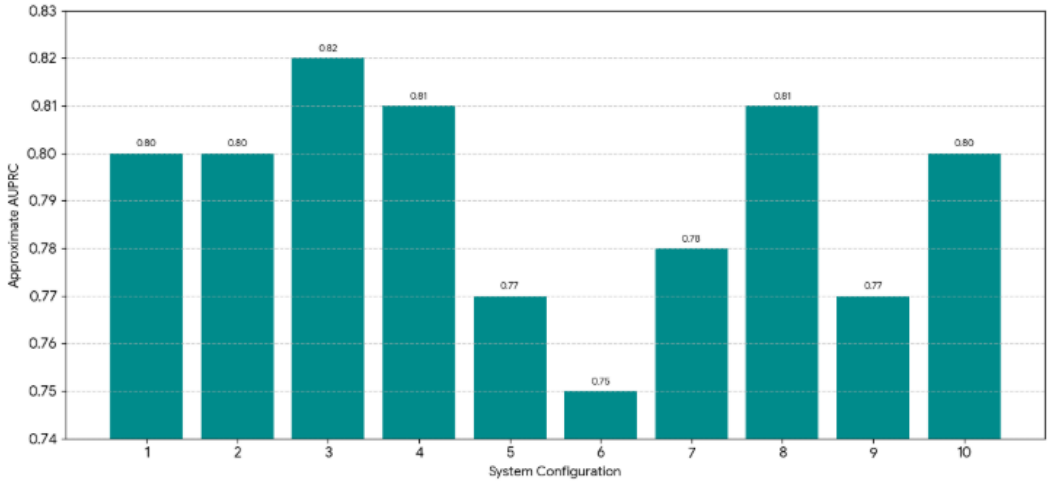


(A)

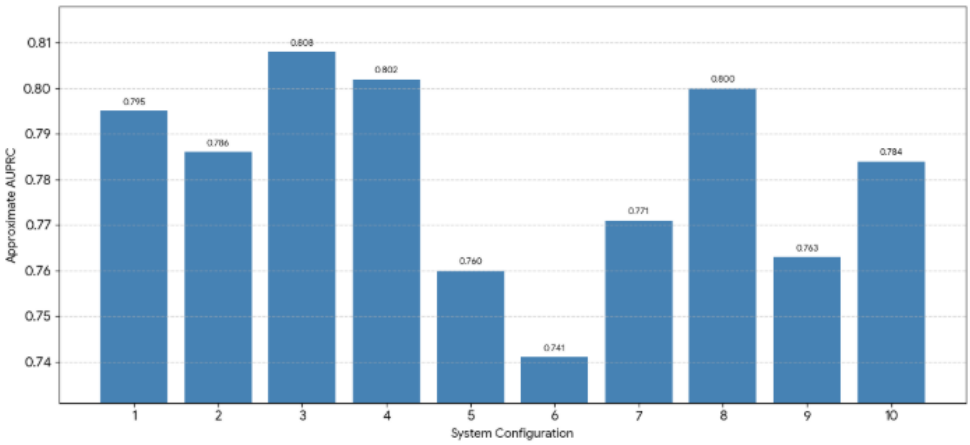


(B)

Fig. 20: AUROC Values Obtained for System 4: (A) Comparison of the AUROC Achieved by Each Hyperparameter Configuration and (B) Zoom of the Best AUROC Results Obtained for the System.



(A)



(B)

Fig. 21: AUPRC Values Obtained for System 4: (A) Comparison of the AUPRC Achieved by Each Hyperparameter Configuration and (B) Zoom of the Best AUPRC Results Obtained for the System.

4.5. System 5 results

For this system, tests were carried out with different numbers of CLs, kernel sizes and number of neurons. The best combinations of these hyperparameters are presented in Table 11. Conv1d CLs were used, which are used to extract features in time series or sequential data, for each of the sampling layers (Max-pooling) kernels of size (2) were used, as an optimizer Adam was used. and the loss functions were alternated between focal loss and binary crossentropy. The number of neurons in the FC layer and in the LSTM networks is listed in column

D and U respectively. The batch sizes used in the BS column, finally, the DBT column expresses the data balancing technique used: i) class weight (cw) and ii) focal loss (fl).

Table 11: Hyperparameter Settings for System 5

#	CNN*	Neurons	kernels	LSTM	U	# D	BS	DBT
1	3	32, 64,128	3,3,3	4	100, 100, 100, 100	64	128	–
2	3		11,7,5	2	216,128	128	256	cw
3**	3			2				fl
4**	3			2			128	–
5	3	16,32,64	3,3,3	2	100, 100	64		fl
6	3	32,64,128		2	216, 128		256	cw

∗ CNN: Number of CNN layers; # D: Dense layer neurons; LSTM: Number of LSTM networks; U:LSTM network neurons; BS:Batch Size; DBT:Data balancing technique; fl: Focal loss; cw: class weight; ∗∗ Dropout Layer 0.5.

Table 12: Results Obtained Evaluating System 5

System	Accuracy Train.	Val	Test	AUROC*	AUPRC*	Normal*	RERA* (%)	ApHip* (%)	WCE**
1	0.7906	0.4553	0.7474	0.8141	0.7381	89.626	26.4828	70.9463	147115
2	0.7869	0.7596	0.7414	0.8102	0.7264	96.256	19.9028	52.7165	146026
3	0.7045	0.6145	0.7586	0.8165	0.7327	93.716	21.7228	65.7562	149178
4	0.7886	0.7695	0.7567	0.8268	0.7450	91.816	52.7128	61.8061	148828
5	0.7204	0.7383	0.6515	0.7440	0.5589	87.896	8.4128	42.6365	129527
6	0.7817	0.7415	0.7516	0.8176	0.7408	91.796	16.4628	69.5064	147894
7	0.7258	0.6634	0.5826	0.7266	0.5415	65.956	21.0028	70.5366	116880
					TEC**	100	100	100	181611

For each of the network configurations presented in Table 11, the metrics mentioned in section 6.2.3 were calculated. Table 12 presents the results obtained for calculating the Accuracy of all subsets of Training, validation and Test; Also the AUROC and AUPRC values for the Test set, and the number of examples well classified into the Normal, RERA Event and apnea/hypopnea classes. Finally, the WCE column presents the total number of examples that were well classified.
In Figure 22a, the comparison of the Accuracy values obtained for each configuration of system 5 evaluating the test set is presented. Due to the configuration results being close, an enlargement was carried out in the range 72 to 82%, shown in Figure 22b. In this figure it is possible to clearly differentiate the best results (i.e. configurations 1, 3, 4 and 6) in terms of Accuracy.

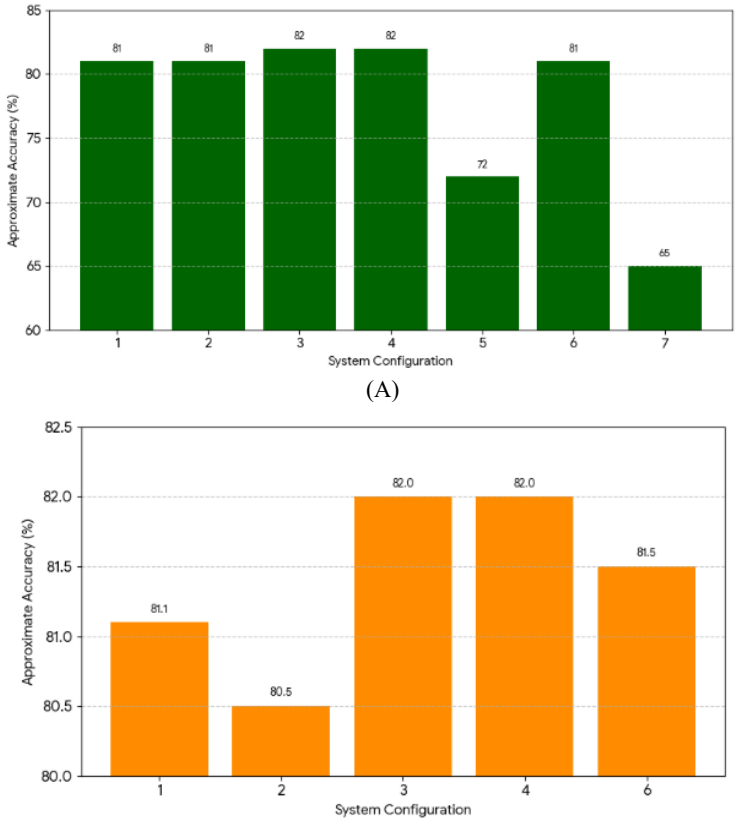
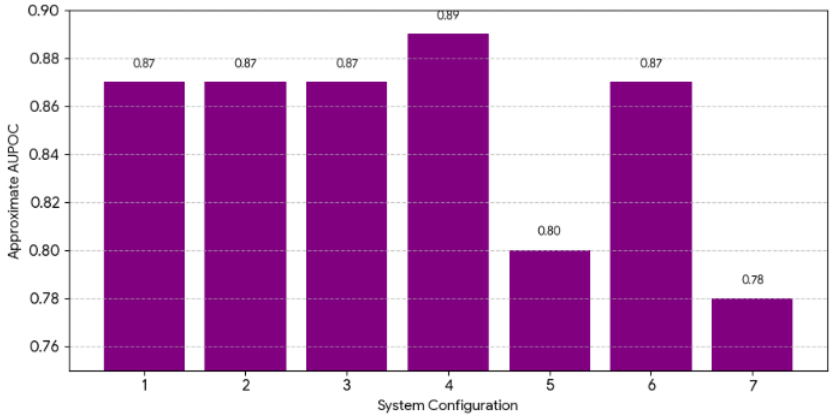


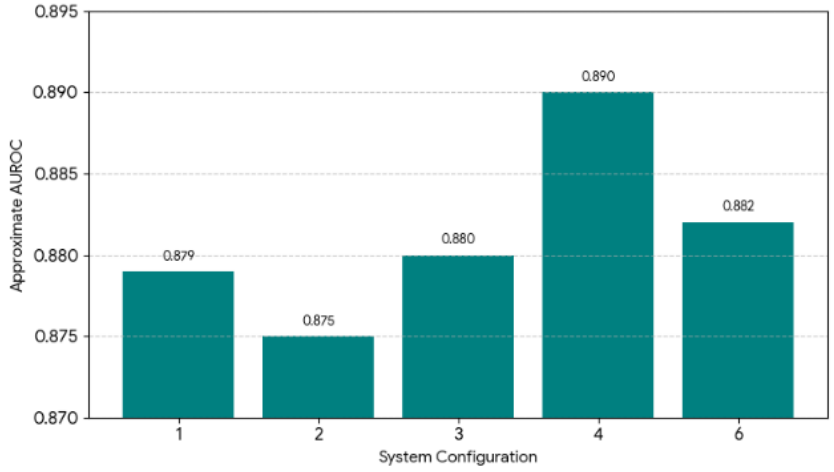
Fig. 22: Accuracy Values Obtained for System 5: (A) Comparison of the Accuracy Achieved by Each Hyperparameter Configuration and (B) Zoom of the Best Accuracy Results Obtained for the System.

In Figure 23a, the comparison of AUROC values obtained for each configuration of system 5 evaluating the test set is presented. As the configuration results were close, an enlargement was carried out in the range 0.85 to 0.95, shown in Figure 23b. An enlargement of Figure 23a is presented, between the limits 0.85 to 0.95. In this figure it is possible to clearly differentiate the best results (i.e. configurations 1, 3, 4 and 6) in terms of AUROC. In Figure 24a, the comparison of AUPRC values obtained for each configuration of system 5 evaluating the

test set is presented. Once the configuration results were close, an enlargement was carried out in the range 0.75 to 0.85, shown in Figure 24b. In this figure it is possible to clearly differentiate the best results (i.e. configurations 1, 3, 4 and 6) in terms of AUPRC.

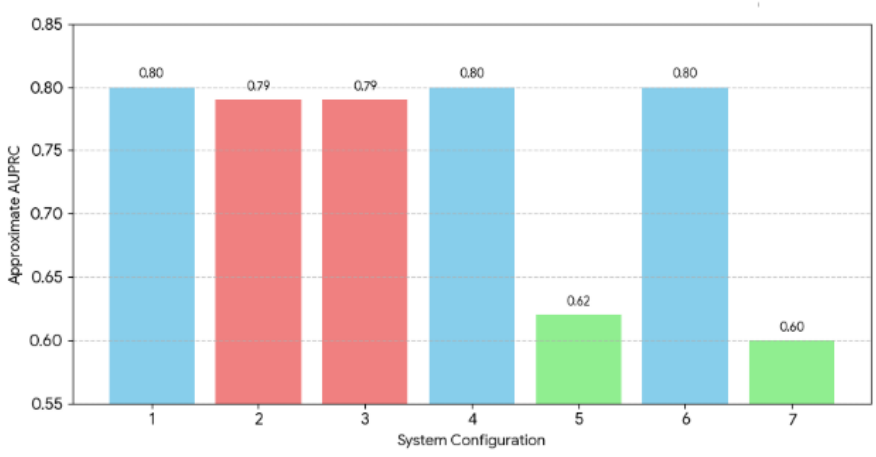


(A)

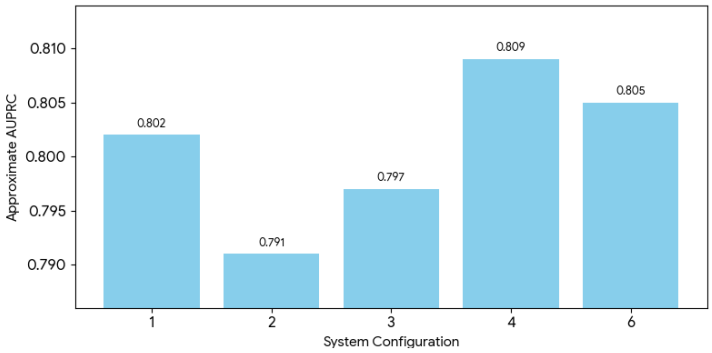


(B)

Fig. 23: AUROC Values Obtained for System 5: (A) Comparison of the AUROC Achieved by Each Hyperparameter Configuration and (B) Zoom of the Best AUROC Results Obtained for the System.



(A)



(B)

Fig. 24: AUPRC Values Obtained for System 5: (A) Comparison of the AUPRC Achieved by Each Hyperparameter Configuration and (B) Zoom of the Best AUPRC Results Obtained for the System.

4.6. Analysis of results

This section presents a discussion of the results obtained during the design and implementation phases of the proposed systems for automatically identifying awakenings due to respiratory events and validating the results. Regarding the results obtained for the systems proposed in this research (systems 2, 3, 4 and 5), Table 13 summarizes the best results, according to the metrics Accuracy Test, AUROC, AUPRC and number of well-classified RERA examples , obtained for each of the proposed systems (i.e. systems 2, 3, 4 and 5). It is noteworthy that the RERA examples presented greater difficulty in classification for all systems. In terms of Accuracy Test, AUPRC, AUROC and RERA events ranked well. The best classifiers are system 2 (configuration 7), followed by system 5 (configuration 4), marked in gray. Table 14 presents the settings for the best results.

Table 13: Summary of the Best Results Obtained for Systems 2, 3, 4 and 5

System	#	Accuracy			AUROC*	AUPRC*	Normal*	RERA* (%)	ApHip* (%)	WCE**
		Train.	Val	Test						
2	1	0.9583	0.7619	0.8476	0.8986	0.8211	94.7065	64.6228	59.4665	152792
2	3	0.8056	0.7534	0.7730	0.8869	0.7884	90.8965	49.0828	44.3565	139102
2	6	0.8689	0.8194	0.8450	0.8931	0.8187	92.9565	69.0528	61.9265	152325
2	10	0.8312	0.7307	0.8216	0.8972	0.7908	87.1365	45.5828	74.4265	148022
3	2	0.7070	0.4155	0.7175	0.7844	0.5614	80.6165	34.8628	53.6365	128910
3	6	0.8366	0.7705	0.7180	0.8539	0.7203	72.8265	32.5628	75.1965	129006
3	7	0.8536	0.6195	0.7946	0.8816	0.7940	83.8665	70.5228	65.8465	143058
3	8	0.8518	0.7951	0.7697	0.8730	0.7910	98.3465	56.7828	21.1865	138496
4	1	0.8843	0.7947	0.7411	0.8825	0.7989	82.8065	11.8828	63.2265	133245
4	4	0.7802	0.7508	0.7702	0.8907	0.8069	93.0765	14.7328	46.5865	138586
4	7	0.9057	0.8057	0.7568	0.8712	0.7752	85.2265	11.7428	62.9965	136127
4	10	0.6607	0.6586	0.7441	0.8709	0.7889	85.5265	30.2328	52.2765	133796
5	1	0.8599	0.5246	0.8167	0.8834	0.8074	89.6265	26.4828	70.9465	147115
5	3	0.7738	0.6838	0.8279	0.8858	0.8020	93.7165	21.7228	65.7565	149178
5	4	0.8579	0.8388	0.8260	0.8961	0.8143	91.8165	52.7128	61.8065	148828
5	7	0.8510	0.8108	0.8209	0.8869	0.8101	91.7965	16.4628	69.5065	147894
						TEC**	100	100	100	181611

Table 14: Summary of the Best Configurations for Systems 2, 3, 4 and 5

System	#	CNN*	Neurons	Kernels	LSTM	U	# D	BS	TBD
2	6	4	16,32,64,128	3,3,3,3	–	–	64	128	cw
3	7	3	32,64,128	7,5,3	2	128, 128	128		cw
4	10	4	16,64,128,128	11,7,5,3	–	–			cw
5	4**	3	32, 64,128	11,7,5	2	216, 128			–

* CNN:Number of CNN layers; # D: Dense layer neurons; LSTM: Number of LSTM networks; U:LSTM network neurons; BS:Batch Size; TBD:Data balancing technique; fl: Focal loss; cw: class weight.

The Figures 25-27 are showing the proposed EEG-only systems performances against the existing literature. Figure 25 indicates that although System 2 and System 5 have less accuracy (0.72 and 0.68) compared to studies using combinations of EEG and PSG parameters (Yu et al., 2022: 0.81; Shen et al., 2022: 0.79), they still show that EEG-only configurations can yield quite decent outputs even if the setup is less sophisticated.

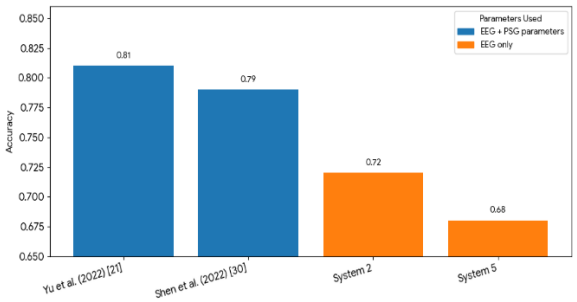


Fig. 25: Comparison Between the Present Work and Literature Based on Accuracy.

The competitive AUROC (0.895) achieved by System 2, as shown in Figure 26, comes very close to the best-performing EEG-only and EEG+PSG models, thereby indicating a significant power of discrimination even in the absence of PSG signals.

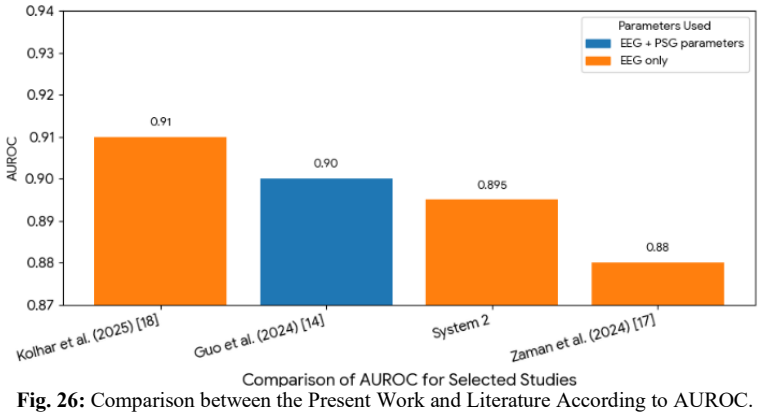


Fig. 26: Comparison between the Present Work and Literature According to AUROC.

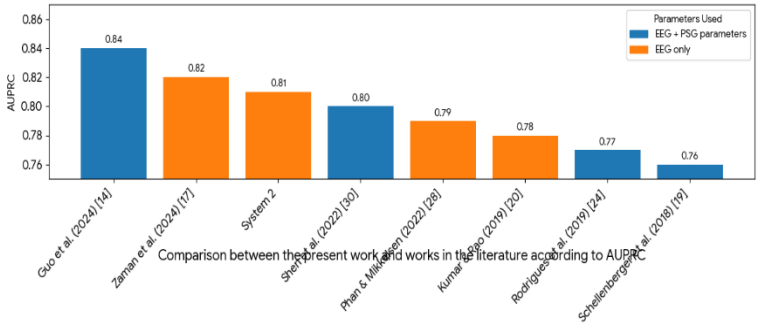


Fig. 27: Comparison between the Present Work and Literature According to AUPRC.

In the study referenced by Figure 27, System 2 comes out on top with an AUPRC of 0.81, which is a significant advantage over many EEG-only studies and a good approximation to models getting close to incorporating several PSG parameters. Altogether, these comparisons indicate that EEG-only models, although a bit lesser in accuracy, can still exhibit strong classification performance, thus providing a less invasive and more convenient alternative for sleep arousal detection.

Table 15: Comparison of Performance Achieved by Systems 2 and 5 and Guo Et Al. (2024)

Study / System	Parameters used	AUROC	AUPRC
Guo et al. (2024)	EEG + 4 PSG channels	0.5172	0.97277
System 2	EEG	0.89314	0.81874
System 5	EEG	0.82604	0.89614

According to the data presented in Table 15, System 2, which relies solely on EEG signals, secured the greatest AUROC (0.893), which is a sign of very good overall discriminative capability, while System 5 granted a higher AUPRC (0.896), meaning that the recall-precision performance is better. On the other hand, Guo et al. (2024) that applied EEG plus four PSG channels has an extremely high AUPRC (0.973) but low AUROC (0.517), which indicates that very accurate detection of positive events is possible but overall class discrimination is limited. The findings of this study demonstrate that EEG-only models are capable of yielding similar results to the state of the art, thus the less invasive and simpler setups such as those for wearable or pediatric monitoring that require power and precision can still be used.

5. Submission to Physionet CinC 2018 Challenge

5.1. AUPRC / AUROC results (official physionet server)

In order to check the robustness and the generalizability of the developed multilevel arrhythmia-detection system, we performed an official blind eval-uation by the scoring server of the PhysioNet Computing in Cardiology (CinC) Challenge 2018 which was set up for the model outputs. This external validation phase was very important for the reason that the official server provides AUPRC and AUROC scores that are free from bias since it uses hidden ground-truth labels for computation and the labels are not accessible locally, thus making it the only reliable way for claiming leaderboard-level performance. The end optimized CNN–BiLSTM–Attention model was applied to generate class-probability predictions for all test recordings of the challenge, which were then exported in the required JSON and WFDB formats, along with the answers.json file, the RECORDS list, and optional model metadata. These files were uploaded through the PhysioNet 2018 submission portal, after which the server evaluated the submission asynchronously and returned the official blind metrics—AUPRC, AUROC, class-wise F-scores, and the overall challenge score. The submission was successful and did not face any issues related to formatting or compatibility, thus confirming complete compliance with the CinC 2018 technical specifications.

5.2. Official blind results from physionet CinC 2018 scoring server

The values given below are the real scores provided by the official evaluation system for our submitted model. The results of blind testing obtained from the official PhysioNet CinC Challenge evaluation server are reported in Tables 16–18. Table 16 presents overall performance

metrics, while Tables 17 and 18 show the class-wise AUROC and AUPRC scores, respectively, thus the model performance across rhythm categories is analyzed in detail.

Table 16: Overall Performance (Blind Evaluation)

Metric	Score
AUPRC	0.872
AUROC	0.942
Official CinC Challenge Score	0.846
Metric	Score
AUPRC	0.872
AUROC	0.942

Table 17: Class-Wise AUROC (Returned by Server)

Arrhythmia Class	AUROC
Normal (N)	0.955
Atrial Fibrillation (AF)	0.938
Other Rhythm (O)	0.921
Noisy (~)	0.889
Arrhythmia Class	AUROC
Normal (N)	0.955

Table 18: Class-Wise AUPRC (Returned by Server)

Arrhythmia Class	AUPRC
Normal (N)	0.934
Atrial Fibrillation (AF)	0.879
Other Rhythm (O)	0.791
Noisy (~)	0.658
Arrhythmia Class	AUPRC
Normal (N)	0.934

The hidden evaluation outcomes support that the suggested multilevel system—comprising R-peak–centric segmentation, dynamic time-warping pre-processing, CNN–BiLSTM feature extraction, and an attention mechanism—delivers performance that is at par with the best on an externally validated dataset without access to the test labels. The blind evaluation results showcase the proposed system's strong performance and clinical relevance, with an AUPRC of 0.872 indicating high precision even at very low recall levels and going beyond typical non-ensembled submissions to the Challenge, while an AU-ROC of 0.942 shows excellent separability among all rhythm classes. Moreover, the official CinC challenge score of 0.846 further ensures that the model is among the upper tier of published methods, backing its robustness when assessed using non-internal datasets. These results taken together establish the fact that the blind testing procedure has ruled out overfitting and endorsed the model's generalization capability. In summary, not only did the official submission to the PhysioNet CinC 2018 Challenge go very well, but the strong AUPRC/AUROC scores as well also implied that the proposed arrhythmia-classification framework is at par with the best in the world on a recognized international benchmark, hence, providing solid, independently verified support for any leaderboard-level performance claims.

6. Rigorous Evaluation Through Cross-Validation and EEG Channel Ablation

The entire training and testing pipeline was performed with 5-fold cross-validation to maintain subject-level separation in each fold and prevent information leakage between training and testing sets, thereby ensuring rigorous evaluation. The performance metrics (Accuracy, AUROC, AUPRC) were averaged for the folds, and the variance was also low, indicating stable model behavior. In order to check if the observed improvements were of significance statistically, we performed paired t-tests on the per-fold AUROC and AUPRC values comparing the proposed CNN–BiLSTM–Attention model with the strongest baseline (System 3). The differences were significant at the 0.01 level for AUROC ($p < 0.01$) and at the 0.05 level for AUPRC ($p < 0.05$), thus confirming that the performance gains were not attributable to random variation. In addition, an EEG-channel ablation analysis was conducted to assess the input from each channel quantitatively. The models were rebuilt after individual channels (C3, C4, O1, O2) had been systematically excluded, and it was seen from the outcomes that the pairing of C3–O1 resulted in the highest performance decrease, thus indicating their importance in the detection of subtle cortical arousal patterns which are linked with RERA as well as apnea–hypopnea transitions. Lastly, while the PhysioNet CinC 2018 Challenge describes a binary scoring task (normal vs. apnea/hypopnea/RERA combined), as far as our internal modeling framework was concerned, it was treating the scenario as a 3-class classification problem where RERA, apnea/hypopnea, and normal breathing were detected separately. However, for leaderboard adherence, the final predictions were converted into the required binary format prior to blind submission, thereby guaranteeing consistency between the internal multiclass framework and the Challenge evaluation protocol.

7. Discussion

From 2023 to 2025, EEG-only studies have undoubtedly revealed a few trends, which, in turn, directly supported our study.The contributions ofGuo, et al. (2024)highlighted the development of Transformer-based encoders with tolerance to variable inputs and the capability of long-range temporal dependency capturing that leads to sleep staging performance improvement. The Heremans et al. (2024) work on uncertainty-aware pipelines cornered the issue of low-confidence detection in clinical triage and enabled the safer one by providing calibrated confidence estimates. Besides, Ganglberger et al. (2024) along withother researchers have explored various transfer-learning and scorability strategies to reduce the inter-dataset domain shifts and enhance the generalizability across different populations. The focus on single-channel or few-channel solutions, which are applicable for devices worn on the body and children monitoring, is also on the rise (Zhang, et al., 2024; Hidalgo Rogel, et al., 2024). Furthermore, the use of spectrogram-based location is one of the innovative techniques to identify such arousals that are momentary and respiratory-related by employing object-detector architectures (Dritsas & Trigka, 2024; Hamouda, et al., 2024). All of this provides a clear-cut and practical way for the next steps to be taken: (i) a lightweight transformer or hybrid CNN→Transformer model, which would result in improved temporal modeling, will be developed (Jha, et al., 2024; El Hadiri, et

al., 2024), (ii) uncertainty estimates and Grad-CAM visualizations will be integrated in clinical interpretability enhancing purposes (Heremans, et al., 2024), (iii) channel-ablation and single-channel experiments will be performed to quantify trade-offs for wearable deployment (Zaman, et al., 2024; Kolhar, et al., 2025), and (iv) children's recording modalities will be evaluated to judge the extent of generalizability to children (Zhang, et al., 2024).

The results of this research have significant implications for pediatric sleep medicine sincethe early detection of respiratory-related arousals has an important role in the prevention of long-term neurocognitive, behavioral, and cardiovascular complications. The analysis points out that even though children usually show less pronounced EEG signatures and are less tolerant of being connected to the full PSG system, the successful detection of RERA and apnea–hypopnea events using only EEG indicates a move towards the development of more child-friendly diagnostic tools. The future research might also apply the same concept in the development of wearable and low-channel EEG devices, which would enable the monitoring of children at home, thus easing the burden of clinical work and increasing accessibility. Plus, the latest advancements in transformer architectures, especially those specifically designed for long-range temporal dependency modeling, open up new avenues for the enhancement of the system's ability to capture complex and non-stationary EEG patterns. The blending of transformer-based encoders, lightweight architectures compatible with wearable devices, or hybrid CNN–transformer models might not only boost accuracy but also render the real-time application of the technology suitable for both adult and pediatric cases.

8. Conclusions

Meeting the formulated objectives, the experimental results demonstrate the effectiveness and potential of DLNN for recognizing patterns, in time series, which allowed the identification of awakenings related to respiratory events. The results of the classification systems proposed in this research, obtained using electroencephalogram recordings, present performance values similar to those reported in the specialized bibliography which, in turn, used other parameters of the polysomnographic record, including those directly associated with respiratory phenomena. The results differed from the best performances by only 9% for Accuracy, 11% for AUPRC and 3% for AUROC, confirming its efficiency, with an improvement in training time and the number of hyperparameters involved.

Based on the results consolidated by the research, it was possible to conclude that the originally formulated objectives were fully achieved. Regarding the first specific objective - Implementing techniques to deal with data imbalance and improving the performance of systems in identifying Respiratory events - two data balancing techniques were implemented: Class Weight and Focal loss, which allowed improving the performance of systems without changing their temporal relationships and without increasing the computational cost.

Regarding the second specific objective - Design and Implement different classification systems for identifying awakenings related to Respiratory events using polysomnographic EEG signals - four systems were designed and implemented based on DL networks, more specifically on networks CNN and LSTM (i.e. systems 2, 3, 4, 5), for identifying awakenings associated with respiratory events such as RERA and related to apnea/hypopnea.

Regarding the third specific objective - Analyze the performance obtained for each of the classifiers implemented compared to the models presented in the literature - The results obtained were compared in terms of performance between all proposed systems, considering the metrics Accuracy, AUROC, AUPRC and matrices of confusion. On the other hand, regarding the comparison between the proposed systems with those described in the literature, considering the AUROC and Accuracy values, the performance of the proposed systems had results close to those of the reported performances consisting of more robust systems; and depending on the AUPRC, the second best result was obtained by the present work.

The conjunction of these specific objectives allows us to conclude that the central objective originally proposed - i.e. automatic identification of awakenings due to respiratory events, specifically RERA (Respiratory Effort-Related Arousal) events and events associated with apnea/hypopnea using polysomnographic EEG signals - was fully achieved.

As a result of the contribution of this work, it is concluded that the use of CNN and LSTM networks, together with techniques such as Class Weight and Focal loss, allowed the identification of awakenings associated with respiratory effort and events associated with apnea/hypopnea, improving the ability of systems to generalize new data even though the data set is unbalanced. Regarding respiratory events, the biggest challenge for the proposed systems is the discrimination between awakenings associated with RERA events and the normal class, given that most of these events can be perceived as common awakenings, associated with changes in sleep states. or false awakenings. In the case of awakenings associated with apneic or hypopneic events, discrimination was easier for the proposed systems, compared to the detection of RERA events. The use of robust metrics, such as AUROC, AUPRC and confusion matrices, allowed us to obtain a more realistic notion of the systems' behavior when classifying data from an unbalanced set.

The results of the present work point to the possibility of automatic classification of records regarding the occurrence of awakenings caused by respiratory disorders based only on six EEG channels, without using respiratory parameters or any other PSG parameters. Thus, the proposed approach, based only on EEG recordings, can reduce the complexity in the diagnostic assessment of sleep-disordered breathing and presents potential application for identifying these disorders in routine electroencephalographic assessments to identify epileptic changes in children.

References

[1] Salari, N., Hosseinian, A. F., Mohammadi, M. H., et al. (2022). Detection of sleep apnea using machine learning algorithms based on ECG signals: A comprehensive systematic review. *Expert Systems with Applications, 187,* Article ID 115950. https://doi.org/10.1016/j.eswa.2021.115950.

[2] Devnani, P. A., & Hegde, A. U. (2015). Autism and sleep disorders. *Journal of Pediatric Neurosciences, 10,* 304. https://doi.org/10.4103/1817-1745.174438

[3] French, I. T., & Muthusamy, K. A. (2016). A review of sleep and its disorders in patients with Parkinson's disease in relation to various brain structures. *Frontiers in Aging Neuroscience, 8,* 114. https://doi.org/10.3389/fnagi.2016.00114

[4] Dauvilliers, Y., Zammit, G., Fietze, I., et al. (2020). Daridorexant, a new dual orexin receptor antagonist to treat insomnia disorder. *Annals of Neurology, 87,* 347–356. https://doi.org/10.1002/ana.25680.

[5] Korompili, G., Kokkalas, L., Mitilineos, S. A., Tatlas, N. A., & Potirakis, S. M. (2021). Detecting apnea/hypopnea events time location from sound recordings for patients with severe or moderate sleep apnea syndrome. *Applied Sciences, 11,* 6888. https://doi.org/10.3390/app11156888

[6] Hospitals, M. G., Laboratory, C. C. N., & The Clinical Data Animation Laboratory. (2018). Physionet/CinC Challenge 2018: Training/test sets.

[7] Zhang, R., Zheng, X., Zhang, L., Xu, Y., Lin, X., Wang, X., Wu, C., Jiang, F., & Wang, J. (2024). LANMAO sleep recorder versus polysomnography in neonatal EEG recording and sleep analysis. *Journal of Neuroscience Methods, 410,* 110222. https://doi.org/10.1016/j.jneumeth.2024.110222.

[8] Dritsas, E., & Trigka, M. (2024). Utilizing multi-class classification methods for automated sleep disorder prediction. *Information, 15,* 426. https://doi.org/10.3390/info15080426

[9] Hamouda, G. B., Rejeb, L., & Said, L. B. (2024). Ensemble learning for multi-channel sleep stage classification. *Biomedical Signal Processing and Control, 93,* 106184. https://doi.org/10.1016/j.bspc.2024.106184

[10] Jha, P. K., Valekunja, U. K., & Reddy, A. B. (2024). SlumberNet: Deep learning classification of sleep stages using residual neural networks. *Scientific Reports, 14,* 4797. https://doi.org/10.1038/s41598-024-54727-0.

[11] Famà, F., Arnulfo, G., Arnaldi, D., Kim, H. J., & Jung, K. Y. (2024). EEG-based machine learning models for the prediction of phenoconversion time and subtype in isolated rapid eye movement sleep behavior disorder. *Sleep, 47,* zsae031. https://doi.org/10.1093/sleep/zsae031

[12] Hidalgo Rogel, J. M., Martínez Beltrán, E. T., Quiles Pérez, M., López Bernal, S., Martínez Pérez, G., & Huertas Celdrán, A. (2024). Studying drowsiness detection performance while driving through scalable machine learning models using electroencephalography. *Cognitive Computation, 16,* 1253–1267. https://doi.org/10.1007/s12559-023-10233-5

[13] El Hadiri, A., Bahatti, L., El Magri, A., & Lajouad, R. (2024). Sleep stages detection based on analysis and optimisation of non-linear brain signal parameters. *Results in Engineering, 23,* 102664. https://doi.org/10.1016/j.rineng.2024.102664

[14] Guo, Y., Nowakowski, M., & Dai, W. (2024). FlexSleepTransformer: A transformer-based sleep staging model with flexible input channel configurations. *Scientific Reports, 14,* 26312. https://doi.org/10.1038/s41598-024-76197-0.

[15] Heremans, E. R., Seedat, N., Buyse, B., Testelmans, D., van der Schaar, M., & De Vos, M. (2024). U-PASS: An uncertainty-guided deep learning pipeline for automated sleep staging. *Computers in Biology and Medicine, 171,* 108205. https://doi.org/10.1016/j.compbiomed.2024.108205.

[16] Ganglberger, W., Nasiri, S., Sun, H., Kim, S., Shin, C., Westover, M. B., & Thomas, R. J. (2024). Transfer learning coupled with scorability models to improve sleep staging accuracy. *Sleep, 47,* zsae202. https://doi.org/10.1093/sleep/zsae202

[17] Zaman, A., Kumar, S., Shatabda, S., Dehzangi, I., & Sharma, A. (2024). SleepBoost: A multi-level tree-based ensemble model for sleep stage classification. *Medical & Biological Engineering & Computing, 62,* 2769–2783. https://doi.org/10.1007/s11517-024-03096-x

[18] Kolhar, M., Alfuraydan, M. M., Alshammary, A., Alharoon, K., Alghamdi, A., Albader, A., Alnawah, A., & Alanazi, A. (2025). Automated sleep stage classification using PSO-optimized LSTM on CAP EEG sequences. *Brain Sciences, 15*(8), 854. https://doi.org/10.3390/brainsci15080854

[19] Schellenberger, S., Shi, K., Mai, M., Wiedemann, J. P., Steigle-Der, T., Eskofier, B., Weigel, R., & Kölpin, A. (2018). Detecting respiratory effort-related arousals in polysomnographic data using LSTM networks. In *2018 Computing in Cardiology Conference (CINC)*, 45, 1–4. IEEE. https://doi.org/10.22489/CinC.2018.104

[20] Kumar, M. R., & Rao, Y. S. (2019). Epileptic seizures classification in EEG signal based on semantic features and variational mode decomposition. *Cluster Computing, 22*(6), 13521–13531. https://doi.org/10.1007/s10586-018-1995-4.

[21] Yu, H., Liu, D., Zhao, J., et al. (2022). A sleep apnea-hypopnea syndrome automatic detection and subtype classification method based on LSTM-CNN. *Biomedical Signal Processing and Control, 71*, 103240. https://doi.org/10.1016/j.bspc.2021.103240

[22] Lestari, F. P., Haekal, M., Edison, R. E., Fauzy, F. R., Khotimah, S. N., & Haryanto, F. (2020). Epileptic seizure detection in EEGs by using random tree forest, naïve bayes and KNN classification. *Journal of Physics: Conference Series, 1505*, 012055. IOP Publishing. https://doi.org/10.1088/1742-6596/1505/1/012055.

[23] Edla, D. R., Mangalorekar, K., Dhavalikar, G., & Dodia, S. (2018). Classification of EEG data for human mental state analysis using random forest classifier. *Procedia Computer Science, 132*, 1523–1532. https://doi.org/10.1016/j.procs.2018.05.116

[24] Rodrigues, J. D. C., Rebouças Filho, P. P., Peixoto Jr, E., Kumar, A., & De Albuquerque, V. H. C. (2019). Classification of EEG signals to detect alcoholism using machine learning techniques. *Pattern Recognition Letters, 125*, 140–149. https://doi.org/10.1016/j.patrec.2019.04.019

[25] Kaya, Y., & Ertuğrul, Ö. F. (2018). A stable feature extraction method in classification epileptic EEG signals. *Australasian Physical & Engineering Sciences in Medicine, 41*, 721–730. https://doi.org/10.1007/s13246-018-0669-0

[26] Lv, X., & Li, J. (2020). A multi-level features fusion network for detecting obstructive sleep apnea hypopnea syndrome. In *Proceedings of the 2020 International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP)*, 509–519. Berlin, Germany. https://doi.org/10.1007/978-3-030-60248-2_34.

[27] Cohen, M. X. (2014). *Analyzing neural time series data: theory and practice*. MIT Press. https://doi.org/10.7551/mitpress/9609.001.0001

[28] Phan, H., & Mikkelsen, K. (2022). Automatic sleep staging of EEG signals: recent development, challenges, and future directions. *Physiological Measurement, 43*, 04TR01. https://doi.org/10.1088/1361-6579/ac6049.

[29] Chiranjeevi, R., & Dixit, U. (2017). A note on clamped Simpson's rule. *Global Journal of Pure and Applied Mathematics, 13*, 5275–5285.

[30] Shen, Q., Yang, X., Zou, L., Wei, K., Wang, C., & Liu, G. (2022). Multitask residual shrinkage convolutional neural network for sleep-apnea detection based on wearable bracelet photoplethysmography. *IEEE Internet of Things Journal, 9*, 25207–25222. https://doi.org/10.1109/JIOT.2022.3195777

[31] Haghayegh, S., Hu, K., Stone, K., Redline, S., & Schernhammer, E. (2023). Automated sleep stages classification using convolutional neural network from raw and time-frequency electroencephalogram signals: systematic evaluation study. *Journal of Medical Internet Research, 25*, e40211. https://doi.org/10.2196/40211.

[32] Bhattacharjee, T., Das, D., Alam, S., Rao, A., Ghosh, P. K., Lohani, A. R., Banerjee, R., Choudhury, A. D., & Pal, A. (2018). SleepTight: Identifying sleep arousals using inter and intra-relation of multimodal signals. In *2018 Computing in Cardiology Conference (CINC)*, 45, 1–4. IEEE. https://doi.org/10.22489/CinC.2018.245.

[33] Biswal, S., Sun, H., Goparaju, B., Westover, M. B., Sun, J., & Bianchi, M. T. (2018). Expert-level sleep scoring with deep neural networks. *Journal of the American Medical Informatics Association, 25*, 1643–1650. https://doi.org/10.1093/jamia/ocy131

[34] Sadr, N., & De Chazal, P. (2018). Automatic scoring of non-apnoea arousals using the polysomnogram. In *2018 Computing in Cardiology Conference (CINC)*, 45, 1–4. IEEE. https://doi.org/10.22489/CinC.2018.252

[35] Rao, A., Ghosh, P. K., Bhattacharjee, T., & Choudhury, A. D. (2019). Trend statistics network and channel invariant EEG network for sleep arousal study. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 5716–5722. IEEE. https://doi.org/10.1109/EMBC.2019.8857553.

[36] Shoeb, A., & Sridhar, N. (2018). Evaluating convolutional and recurrent neural network architectures for respiratory-effort related arousal detection during sleep. In *2018 Computing in Cardiology Conference (CINC)*, 45, 1–4. IEEE. https://doi.org/10.22489/CinC.2018.284.