

Multimodal Deep Learning and Feature Fusion for Automated Sleep Apnea Event Detection Using ECG and Respiratory Signals

Khushi¹
M. Tech, ECE

Vikas Sindhu²
Associate Professor, ECE

Department of Electronics and Communications Engineering,
University Institute of Engineering and Technology (UIET), M.D. University,
Rohtak – 124001(Haryana)

Abstract - Accurate and automated assessment of cardiorespiratory function is essential for physiological monitoring and the early detection of sleep-related disorders. Sleep apnea is a common but frequently underdiagnosed condition, affecting approximately 10% of middle-aged individuals, while polysomnography (PSG), the current diagnostic gold standard, is resource-intensive and requires substantial manual scoring. Advances in biosignal acquisition using modalities such as electrocardiography (ECG), photoplethysmography (PPG), electromyography (EMG), vibromyography (VMG), mechanomyography (MMG), and respiratory signals have enabled high-resolution and multimodal physiological analysis. In this context, deep learning techniques, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and hybrid CNN-LSTM architectures, offer automated feature extraction and the ability to capture complex spatial and temporal patterns in physiological signals. This study investigates a multimodal deep-learning framework for automated sleep-apnea event detection by combining learned representations from raw ECG and respiratory signals with engineered cardiorespiratory features, including heart-rate-variability (HRV) and respiration-related descriptors. Using the MIT-BIH Polysomnographic Database, 10,274 non-overlapping 30-second windows from 18 records were analyzed. To prevent record-level data leakage, the records were divided into locked training, validation, and test sets. A logistic-regression model based on engineered features, a one-dimensional CNN based on raw physiological signals, and a CNN-based feature-fusion model were evaluated. The multimodal fusion model achieved a test ROC-AUC of 0.7341, PR-AUC of 0.5778, accuracy of 0.7317, precision of 0.5689, recall of 0.7204, specificity of 0.7372, and F1-score of 0.6358. Compared with logistic regression, the fusion approach improved ROC-AUC by 0.0222 and F1-score by 0.0665, while outperforming the raw-signal CNN by 0.0833 and 0.1440, respectively. These findings indicate that combining learned raw-signal representations with compact engineered cardiorespiratory descriptors can improve sleep-apnea event classification compared with either strategy alone. Future extensions could incorporate hybrid CNN-LSTM architectures for improved temporal modeling, Transformer-based approaches for multimodal signal fusion, and diffusion-based generative models for realistic biosignal augmentation and robustness evaluation. Nevertheless, substantial variability across held-out records highlights the need for external multi-center validation and patient-level statistical evaluation before clinical deployment.

Keywords: *sleep apnea; polysomnography; ECG; respiration; heart-rate variability; convolutional neural network; multimodal fusion; machine learning; record-level validation*

1. INTRODUCTION

Sleep-disordered breathing is characterized by recurrent abnormalities in ventilation during sleep and is associated with important cardiovascular and metabolic consequences. Polysomnography (PSG) remains the reference diagnostic approach, but acquisition and interpretation involve multiple physiological channels and expert scoring. Recent reviews describe growing interest in machine learning and deep learning systems that use ECG, respiratory signals, oxygen saturation, EEG, and combinations of physiological modalities to automate detection and screening [1], [2], [3], [4].

The MIT-BIH Polysomnographic Database is a useful benchmark for this problem because it contains multi-channel overnight recordings with beat-level ECG annotations and sleep-stage/apnea annotations. The database contains 18 records collected from 16 subjects and more than 80 hours of polysomnographic data [5]. The original annotations are organized in 30-second intervals, making the resource particularly suitable for segment-level experiments [6].

A methodological challenge in physiological-signal machine learning is that neighboring windows from the same individual are highly correlated. A random window-level split can therefore produce optimistic estimates because windows from the same recording can occur in both training and test data. The present study addresses this issue by using a locked record-level split throughout model development and final evaluation.

A second challenge is representation. End-to-end CNNs can learn local waveform patterns directly from raw signals, while engineered HRV and respiratory descriptors provide compact summaries that encode physiologically interpretable information. The central hypothesis of this work is that these representations are complementary: a learned representation can capture morphology and temporal structure, whereas engineered features can expose higher-level cardiorespiratory behavior.

The primary objective was therefore to compare three approaches under the same locked record-level protocol: (1) logistic regression using engineered features, (2) a one-dimensional CNN using raw physiological signals, and (3) a CNN with engineered-feature fusion. The secondary objectives were to quantify uncertainty using bootstrap confidence intervals and to examine record-level variability in the final model.

2. BACKGROUND

2.1 Physiological basis

Apneic and hypopneic events alter ventilation and can produce changes in respiratory amplitude, respiratory timing, autonomic activity, and heart-rate dynamics. ECG-derived information can therefore contain indirect information about breathing and sleep state. Prior work has demonstrated that ECG and respiratory signals can support automated sleep analysis, and systematic reviews have identified ECG and respiratory sensing as major directions in automated sleep-apnea research [2], [3], [7].

2.2 ECG and HRV

The RR interval series provides a compact representation of beat-to-beat cardiac timing. HRV measures such as mean RR, mean heart rate, SDNN, RMSSD, minimum RR, and maximum RR summarize different aspects of cardiac dynamics. In the present study, these descriptors were computed per 30-second window and combined with respiratory statistics.

2.3 Respiratory representation

Respiratory signals contain amplitude and timing information. The engineered representation used in this study included mean, standard deviation, minimum, maximum, range, RMS, energy, and estimated respiratory rate. These descriptors were designed to provide a compact summary rather than replace the raw respiratory waveform.

2.4 Deep learning for physiological signals

CNNs are well suited to one-dimensional physiological signals because convolutional filters can learn local patterns while pooling reduces temporal dimensionality. Previous work has shown that deep networks can learn useful representations from ECG and respiration for sleep-related tasks [7], [8]. Recent systematic reviews also indicate that deep-learning approaches are increasingly used for sleep-apnea detection, although methodological heterogeneity and validation strategy remain important limitations [1], [2], [3], [4].

3. RELATED WORK

The literature contains three broad methodological families. The first uses engineered physiological features with conventional machine learning. These systems are attractive because the features are compact and interpretable, but their performance depends on the adequacy of the handcrafted representation. The second family uses end-to-end deep learning on raw or minimally processed signals. These models can learn task-specific representations but generally require careful regularization and substantial data. The third family combines multiple modalities or learned and engineered representations.

Systematic reviews have reported promising performance for ECG-based machine-learning approaches, while also emphasizing substantial variation in study design, data segmentation, and validation [1]. A broader review of physiological-signal deep learning similarly identified ECG, airflow/respiration, oxygen saturation, and other sensors as common inputs and highlighted window size, signal type, model structure, and database choice as influential methodological factors [2].

Work on ECG and respiration has also demonstrated that combining physiological streams can improve sleep-related inference. For example, deep networks using ECG and respiratory signals have been studied for sleep staging [8]. These findings motivate the multimodal design used here, although sleep staging and apnea-event classification are distinct tasks and should not be conflated.

A recent systematic review and meta-analysis of ECG-based sleep-apnea detection reported high pooled diagnostic performance across a large heterogeneous literature, while also identifying variation and methodological bias as important considerations [1]. The present study differs from many high-performing reports by emphasizing a strict record-level split and transparent uncertainty analysis rather than maximizing a single point estimate under a random-window protocol.

4. MATERIALS AND METHODS

4.1 Study design

This was a retrospective computational study using a publicly accessible physiological-signal database. The analysis was designed as a supervised binary classification task at the 30-second window level. The final manuscript follows the general reporting principles emphasized by TRIPOD+AI for transparent reporting of prediction models using regression and machine-learning methods [9].

4.2 Dataset

The MIT-BIH Polysomnographic Database contains 18 records from sleep-laboratory subjects and more than 80 hours of multi-channel physiological recordings. The database includes ECG, EEG, respiratory and other channels depending on the record, with beat annotations and sleep-stage/apnea annotations [5,6]. The present pipeline used the available ECG/respiratory information and constructed 10,274 30-second windows.

4.3 Windowing and labels

Signals were represented as 30-second windows. At the database sampling rate of 250 Hz, each window contained 7,500 samples per channel. The final raw representation therefore had shape (10,274, 7,500, 2). Labels were binary, with 3,340 positive windows and 6,934 negative windows (32.51% positive).

4.4 Engineered features

Fourteen engineered features were used: mean_RR, mean_HR, SDNN, RMSSD, min_RR, max_RR, resp_mean, resp_std, resp_min, resp_max, resp_range, resp_rms, resp_energy, and resp_rate_bpm. These features were selected as compact cardiac and respiratory descriptors and were not intended to represent an exhaustive physiological feature set.

4.5 Quality control and missing data

The final multimodal dataset contained 10,274 rows with no duplicate record/window keys. Most windows were classified as good quality; 7 were moderate and 2 poor. These windows were retained rather than selectively removed. Five windows had missing respiratory-rate values. No global imputation was performed.

The locked training set contained four missing respiratory-rate values and the test set contained one. Missing respiratory rate was imputed using the training-set median (14.0), ensuring that information from validation and test sets did not influence preprocessing.

4.6 Record-level data partitioning

The final partition was fixed before model evaluation. Training records were slp01b, slp02a, slp14, slp16, slp37, slp41, slp45, slp48, slp59, slp60, slp61, and slp67x. Validation records were slp01a, slp02b, and slp66. Test records were slp03, slp04, and slp32. This produced 7,244 training windows, 950 validation windows, and 2,080 test windows. The positive-window proportions were 32.52%, 32.42%, and 32.50%, respectively.

No record appeared in more than one split, and all 10,274 windows were assigned exactly once. This design was maintained for every subsequent model and evaluation step.

4.7 Preprocessing

For engineered features, missing values were imputed using training-set statistics and all features were standardized with a Standard Scaler fit only on the training data. For raw signals, each channel was normalized using training-set means and standard deviations. Validation and test signals were transformed using these fixed training statistics.

4.8 Baseline model

Logistic regression was trained using the 14 standardized engineered features. Class weighting was used to account for the binary class distribution. Random forest and histogram-based gradient boosting were also evaluated as preliminary baselines; logistic regression was selected because it achieved the highest validation ROC-AUC among the baseline models.

4.9 Raw-signal CNN

The raw-signal CNN received two channels with 7,500 samples per window. The architecture used four one-dimensional convolutional blocks with channel widths 32, 64, 128, and 256, followed by adaptive average pooling and a fully connected classifier. Batch normalization, ReLU activations, max pooling, and dropout were used for regularization. Training used class-weighted binary cross-entropy and early stopping based on validation ROC-AUC.

4.10 CNN-feature fusion

The final model used the same CNN backbone to generate a 64-dimensional learned embedding. In parallel, the 14 engineered features were passed through a small multilayer perceptron producing a 16-dimensional engineered-feature embedding. The two embeddings were concatenated into an 80-dimensional fused representation and passed to a classifier with dropout.

4.11 Threshold selection

The final model's decision threshold was not selected using the test set. Thresholds were evaluated on validation data and the threshold yielding the highest validation F1-score was selected. The resulting threshold was 0.45 and was then fixed before final test evaluation.

4.12 Statistical analysis

Final test performance was reported using ROC-AUC, PR-AUC, accuracy, precision, recall/sensitivity, specificity, negative predictive value, and F1-score. Uncertainty was quantified with percentile bootstrap confidence intervals using 2,000 resamples and a fixed random seed of 42. The bootstrap analysis resampled test windows, so the confidence intervals should be interpreted as window-level uncertainty rather than independent-subject uncertainty.

Paired statistical significance testing between the three models was not performed because raw prediction vectors for the baseline and raw-CNN models were not retained in the original evaluation artifacts. Aggregate metrics alone were not treated as sufficient evidence for a paired significance test.

5. RESULTS

5.1 Dataset

The complete dataset contained 10,274 windows from 18 records. The locked training, validation, and test sets contained 7,244, 950, and 2,080 windows, respectively. Class balance was closely matched across splits.

Split	Windows	Records	Positive	Negative	Positive (%)
Train	7244	12	2356	4888	32.52
Validation	950	3	308	642	32.42
Test	2080	3	676	1404	32.50

5.2 Baseline comparison

Among the engineered-feature baselines, logistic regression achieved the highest validation ROC-AUC (0.7503) and was therefore selected as the baseline comparator. On the locked test set it achieved ROC-AUC 0.7119, accuracy 0.6937, precision 0.5243, recall 0.6228, specificity 0.7279, and F1 0.5693.

5.3 Raw CNN

The raw CNN achieved a best validation ROC-AUC of 0.6721. With the validation-selected threshold of 0.15, its test ROC-AUC was 0.6508, accuracy 0.6293, precision 0.4435, recall 0.5518, specificity 0.6667, and F1 0.4918.

5.4 Fusion model

The CNN-plus-feature-fusion model achieved its best validation ROC-AUC of 0.7936 at epoch 3. The validation-selected threshold was 0.45. On the locked test set, the model achieved ROC-AUC 0.7341, PR-AUC 0.5778, accuracy 0.7317, precision 0.5689, recall 0.7204, specificity 0.7372, NPV 0.8456, and F1 0.6358.

Model	ROC-AUC	PR-AUC	Accuracy	Precision	Recall	Specificity	F1
Logistic regression	0.7119	—	0.6937	0.5243	0.6228	0.7279	0.5693
Raw CNN	0.6508	0.4830	0.6293	0.4435	0.5518	0.6667	0.4918
CNN + feature fusion	0.7341	0.5778	0.7317	0.5689	0.7204	0.7372	0.6358

5.5 Confidence intervals

Bootstrap analysis provided a 95% confidence interval of 0.7079–0.7598 for ROC-AUC and 0.6070–0.6645 for F1. The corresponding intervals for PR-AUC, accuracy, precision, recall, and specificity were 0.5412–0.6204, 0.7135–0.7505, 0.5362–0.6045, 0.6873–0.7548, and 0.7142–0.7592, respectively.

5.6 Confusion matrix and error profile

At the fixed threshold of 0.45, the final model produced 1,035 true negatives, 369 false positives, 189 false negatives, and 487 true positives. There were 558 total classification errors: 369 false positives and 189 false negatives.

5.7 Record-level performance

Performance varied markedly across the three held-out records. For slp03, the fusion model achieved ROC-AUC 0.7800 and F1 0.6629. For slp04, ROC-AUC was 0.8220 and F1 0.7075, with recall of 0.9842. For slp32, ROC-AUC was 0.3070 and the selected threshold resulted in zero positive predictions, producing zero recall and zero F1. This variability is an important finding rather than a result to be hidden by aggregate metrics.

6. LIMITATIONS

1. The database is small at the subject/record level: the final test set contains only three records. Therefore, confidence intervals based on window resampling do not fully capture between-subject uncertainty.
2. The dataset is historical and originates from a single sleep-laboratory resource. External multi-center validation is necessary to assess generalization.
3. The task definition and preprocessing pipeline are specific to the current annotation/windowing design and should not be assumed to correspond directly to a clinical apnea-hypopnea index or patient-level diagnosis.
4. The raw CNN and fusion models were trained on CPU with a relatively compact architecture; broader hyperparameter optimization could potentially improve performance, but it would need to be performed without violating the locked test protocol.
5. Paired statistical significance testing between models was not possible because raw probability vectors for the earlier models were not retained.
6. The final threshold was selected to maximize validation F1 and therefore represents an operating point optimized for this research objective, not a clinically validated decision threshold.
7. The study did not evaluate calibration, decision-curve utility, fairness across demographic subgroups, or prospective clinical workflow integration.

7. CONCLUSION

This study evaluated three approaches for 30-second physiological-window classification using the MIT-BIH Polysomnographic Database: engineered-feature logistic regression, a raw-signal CNN, and a CNN with engineered-feature fusion. Under a locked record-level evaluation protocol, the fusion model achieved the strongest aggregate test performance, with ROC-AUC 0.7341, PR-AUC 0.5778, accuracy 0.7317, recall 0.7204, specificity 0.7372, and F1 0.6358. Bootstrap confidence intervals supported the stability of these window-level estimates.

The central contribution is methodological as much as numerical: combining learned representations with explicit cardiorespiratory descriptors provided a more effective representation than either component alone in this experiment, while record-level evaluation exposed substantial heterogeneity that would be obscured by aggregate metrics alone.

REFERENCES

- [1] M. E. A. 2. O. E. T. 3. Y. R. Y. 4. E. E. O. 5. M. B. Y. 6. Mustafa Eray Kilic 1, "Diagnostic accuracy of machine learning algorithms in electrocardiogram-based sleep apnea detection: A systematic review and meta-analysis," *National Library of Medicine*, vol. 102097., no. doi: 10.1016/j.smr.2025.102097, 2025 Jun:81.
- [2] D. A. 1. Praveen Kumar Tyagi 1, "Systematic review of automated sleep apnea detection based on physiological signal data using deep learning algorithm: a meta-analysis approach," *National Library of Medicine*, Vols. 13(3):293-312. , no. doi: 10.1007/s13534-023-00297-5., 2023 Jul 5.
- [3] Z. D. 1. 2. J. Z. 2. 3. S. Z. 1. 2. Q. W. 2. 3. K. Z. 2. 3. X. W. 1. 3. L. C. 1. 3. Liqing Yang 1 2, "Algorithmic detection of sleep-disordered breathing using respiratory signals: a systematic review," *National Library of Medicine*, vol. 45(3), no. DOI: 10.1088/1361-6579/ad2c13, 2024 Mar 21.
- [4] L. S. 2. Q. Z. 3. Y. Z. 4. 5. Dandan Peng 1, "AI-driven approaches for automatic detection of sleep apnea/hypopnea based on human physiological signals: a review," *National Library of Medicine*, vol. 13(1):7, no. doi: 10.1007/s13755-024-00320-8., 2024 Dec 20.
- [5] G. Moody, "MIT-BIH Polysomnographic Database," Physionet, Aug. 3, 1999.
- [6] G. B. M. Y Ichimaru 1, "Development of the polysomnographic database on CD-ROM," *National Library of Medicine*, Vols. 53(2):175-7, no. 10.1046/j.1440-1819.1999.00527.x, 1999 Apr.
- [7] Q. L. 2. C. L. 3. S. P. S. 4. S. N. 1. G. D. C. 1. 5. 6. Qiao Li 1, "Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram," *National Library of Medicine*, vol. 39(12):124005., no. 10.1088/1361-6579/aaf339., 2018 Dec 21.
- [8] W. G. 1. E. P. 1. M. J. L. 1. S. A. Q. 1. B. G. 1. R. A. T. 1. O. A. 2. R. J. T. 3. M. B. W. 1. Haoqi Sun 1, "Sleep staging from electrocardiography and respiration with deep learning," *National Library of Medicine*, vol. 43(7):zsz306, no. 10.1093/sleep/zsz306., 2020 Jul 13.
- [9] M. K. D. P. e. a. Collins GS, "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods - April 18, 2024," *BMJ.*, vol. 385:e078378, no. doi: <https://doi.org/10.1136/bmj-2023-078378>, 16 April 2024.
- [10] Sun C, et al. A systematic review of deep learning methods for modeling electrocardiograms during sleep. *Physiological Measurement*. 2022;43(8). doi:10.1088/1361-6579/ac826e.