NeuralApnea Triage: Machine Learning Powered ECG Analysis System for Sleep Apnea Detection

Oliver Olejar
Western University
oolejar@uwo.ca

Daniel Kaminsky
Western University
dkamins7@uwo.ca

Annie Liu
Western University
yliu5349@uwo.ca

John MacPhie
Western University
jmacphi2@uwo.ca

Sneha Shah
Western University
sshah495@uwo.ca

Noah Kostesku
Western University
nkostes@uwo.ca

***Abstract*—Sleep apnea is a respiratory disorder marked by recurrent airflow reductions during sleep, leading to intermittent hypoxia and autonomic instability that produce measurable changes in the electrocardiogram (ECG). Polysomnography (PSG), the clinical gold standard for sleep apnea diagnosis, is resource-intensive and often associated with long wait times. As a result, there is strong interest in ECG-based triage tools that can screen large populations and flag higher-risk patients for confirmatory PSG. Using the PhysioNet Apnea-ECG dataset, we evaluate minute-level apnea detection. We train a CNN-Transformer with 5-fold cross-validation and select a checkpoint by validation loss. On the held-out test cohort, the selected model achieves 89.21% accuracy, 84.24% sensitivity, 92.30% specificity, 85.68% F1-score, and an AUROC of 95.60%. We provide Grad-CAM overlays and an explainability agent that summarizes high-confidence minutes for clinician review. Performance is competitive with recent Apnea-ECG studies using the same protocol; external validation is required before clinical deployment. Code repository: https://github.com/Western-Artificial-Intelligence/ecg-classifier-vrllm**

I. Introduction

A. Motivation

Obstructive sleep apnea (OSA) is a prevalent sleep disorder characterized by repeated episodes of upper airway collapse during sleep, leading to intermittent hypoxia and sleep fragmentation that contribute to significant morbidity [1]. The disorder is strongly associated with serious health complications, including cardiovascular disease, hypertension, stroke, and metabolic dysregulation, as well as excessive daytime sleepiness and reduced quality of life [1] [2]. Despite this burden, hundreds of millions of adults are estimated to be affected globally, with roughly 80% of cases remaining undiagnosed and therefore untreated [1]. The current diagnostic gold standard, polysomnography, is expensive, time-consuming, and resource-intensive, with costs ranging from $1,000 to $10,000 per study and requiring specialized equipment and trained technologists [3]. As a result, many regions report polysomnography wait times ranging from several months to over two years, during which patients remain at elevated risk from untreated OSA [4]. Manual scoring of sleep studies requires 1.5 to 2 hours per recording and is subject to inter-rater variability, further limiting throughput and scalability [5]. These constraints create a diagnostic bottleneck that contributes directly to persistent underdiagnosis. In low-resource settings where full polysomnography is unavailable or impractical, there is a particular need for alternative approaches that can function as screening or triage tools rather than full diagnostic replacements [1]. Electrocardiography (ECG) offers a promising modality in this context, as it is inexpensive, widely available, and routinely recorded overnight in cardiac monitoring workflows. Overnight ECG data are frequently collected during Holter monitoring and telemetry, and prior work has demonstrated that apnea-related information can be extracted from these recordings using automated analysis [6]. Large archives of such ECG recordings, including public repositories, therefore represent an underutilized resource that could support scalable OSA screening without new data acquisition. Historically, ECG-based OSA detection has relied on hand-crafted features and heuristic rules, which may fail to capture the complex, nonlinear patterns that characterize autonomic and morphological changes during apnea. Recent deep learning approaches, including convolutional and transformer-based architectures, have shown improved performance in modeling subtle ECG dynamics and heart-rate–related variability compared with traditional feature-based methods [7] [8]. The present work investigates whether a deep learning model operating on single-lead ECG can serve as a scalable screening or triage tool to prioritize patients for confirmatory polysomnography, rather than replace comprehensive sleep studies outright [8] [9].

B. Problem Definition

The objective of this study is to develop and evaluate a software-only screening system that classifies sleep apnea versus normal breathing using single-lead overnight ECG recordings. The system processes multi-minute ECG segments to produce per-segment apnea probabilities, which are then aggregated into a night-level apnea burden or risk score, enabling deployment on existing ECG devices and archived recordings without additional sensors or specialized staff [10]. The classification task is addressed using a hybrid CNN–Transformer architecture. The convolutional neural network (CNN) component extracts local ECG features, including beat morphology and beat-to-beat interval structure, from short time windows. The Transformer encoder then applies multi-head self-attention to these learned feature sequences to model longer-range temporal dependencies spanning several minutes, allowing the model to relate distant segments within the same overnight recording [10]. In line with prior work on apnea detection and ECG time-series modeling, this design aims to better capture apnea-related patterns that evolve gradually over time, addressing limitations of representative methods that either rely on multiple physiological signals or operate on short, weakly contextual segments [11]. The model is trained and evaluated on the PhysioNet Apnea-ECG database, which provides 70 single-lead ECG recordings with expert-derived per-minute labels, to assess generalization to new patients [12]. To mitigate the black-box nature commonly associated with deep learning in medical applications, the system incorporates an explainability agent that operates on per-window model outputs and derived physiological metrics, such as heart rate variability and R-peak amplitude-highlighting windows and ECG regions that appear most salient or influential for the model's predictions [13]. Techniques such as Grad-CAM–style relevance mapping and attention visualization are used to provide clinicians with interpretable views of which parts of the signal the model focuses on, without asserting strict causal attribution. Overall, the system is designed as a software-only screening and triage tool that can run on existing ECG hardware or retrospective datasets, supporting the screening of larger populations, prioritization of higher-risk patients for confirmatory polysomnography, and potential reduction of diagnostic delays in resource-constrained settings [7] [8] [9].

C. Related Works

Zhang et al. (2020) developed an automated multi-model deep neural network for sleep apnea detection using multichannel polysomnography signals, including EEG (C3/A2, C4/A1), EOG (two channels), and EMG, achieving 89.3% detection accuracy across 1,200 subjects. [14]. The architecture combines convolutional layers for spatial feature extraction with recurrent layers for temporal modeling across the multiple physiological channels [14]. Because this approach depends on comprehensive PSG equipment with EEG, EOG, and EMG sensors, its deployment is naturally oriented toward specialized sleep laboratories. By contrast, the present study focuses on single-lead ECG data, with the goal of enabling screening on standard cardiac monitoring devices and retrospective analysis of existing ECG recordings without requiring additional PSG sensors or infrastructure.

Rizal et al. (2022) proposed an obstructive sleep apnea classification method based on heart-rate variability (HRV) features extracted from single-lead ECG signals in the PhysioNet Apnea-ECG database, using one-minute segments annotated as normal or OSA. They computed eleven time-domain HRV parameters from the RR-interval series and trained support vector machines (SVMs) with various kernels, achieving a maximum accuracy of 89.5% using a fine Gaussian SVM over 16,829 one-minute ECG segments. [15]. However, because each one-minute segment is reduced to a static HRV feature vector and classified independently, this approach does not explicitly model longer-range temporal dependencies or evolving apnea patterns across the entire overnight recording. The present study addresses this limitation by operating directly on sequential RR-interval and R-peak amplitude series within multi-minute context windows and applying a hybrid CNN–Transformer architecture, allowing the model to capture both local ECG morphology and broader temporal structure that may be informative for apnea vs non-apnea classification.

Chaw et al. (2019) developed a deep learning–based sleep apnea detection system using nocturnal SpO_2 recordings from 50 subjects, comparing a 10-layer convolutional neural network (CNN) against traditional classifiers such as linear discriminant analysis, support vector machines, and shallow artificial neural networks. Their CNN model, trained on segmented SpO_2 sequences to classify apnea versus non-apnea, achieved an overall accuracy of approximately 91.3%, outperforming the baseline methods while treating the network as a high-capacity feature extractor and classifier for oxygen saturation dynamics [16]. However, the authors characterize the CNN as effectively a black-box model whose internal decision process is not transparent and do not incorporate any explicit explainability mechanism to show which temporal patterns or signal regions drive apnea predictions. In contrast, the present study not only applies a deep neural architecture (CNN–Transformer) to ECG-based apnea screening but also introduces an explainability agent and Grad-CAM–style visualizations to highlight apnea-like segments, aiming to make model outputs more interpretable for clinicians rather than focused primarily on predictive performance.

II. Methodology

A. Dataset

For this study, we used the Apnea-ECG Database [17] released as part of the PhysioNet/Computing in Cardiology Challenge [18]. The dataset contains 70 full-night single-lead ECG recordings, each recording spanning between 7 to 10 hours. All recordings contain annotations from a field expert indicating whether an apnea event occurred during each 60-second interval [17]. These annotations label each segment as either apnea or normal breathing, enabling supervised learning for apnea detection.

Each recording includes several file types. The .dat files contain the digitized ECG waveform (16-bit samples, sampled at 100 Hz, approximately 200 A/D units per millivolt) [17]. The accompanying .hea files are human-readable header files specifying metadata such as signal names, units, calibration parameters, and the format of the associated data file [17]. The .apn files are binary annotation files providing the apnea labels for each minute of the recording [17]. The dataset also provides .qrs files, which contain machine-generated

QRS complex annotations produced by an automated detector. However, these annotations are known to contain substantial errors and inconsistencies. Because they are neither required nor reliable for our analysis, we excluded the .qrs files from our processing pipeline.

Overall, the Apnea-ECG dataset provides a rich, well-labeled, and widely used benchmark for evaluating deep learning models on physiological time-series data related to sleep apnea.

B. Data Preprocessing

For each subject, we loaded the raw single-lead ECG waveform using `wfdb.rdrecord` at the native 100 Hz sampling rate (no resampling), producing a one-dimensional overnight time series (typically >1M samples per record) suitable for downstream QRS morphology and heart rate variability (HRV) analysis [19]. Because reference annotations are provided at 1-minute resolution, we treated each minute as a labeled instance [19]. To add temporal context, we constructed a centered 5-minute window for each target minute (2 minutes before, the target minute, and 2 minutes after), producing 300-second segments (30,000 samples) that capture slower apnea-related heart rate dynamics [20]. Windows were generated with a 1-minute stride, the first and last 2 minutes of each record were excluded due to missing context, and each window inherited the label of its central minute.

We applied a 3 to 45 Hz FIR band-pass filter to reduce baseline wander and high-frequency noise while preserving QRS content, consistent with prior ECG preprocessing for R-peak extraction [21]. Filtering was performed with a zero-phase forward and reverse implementation to avoid phase distortion. R-peaks were detected using the Hamilton QRS algorithm [22] and refined by aligning each detection to the local waveform maximum. We removed windows with clear detection failures or implausible heart rates (below 20 bpm or above 300 bpm), as well as segments dominated by artifacts or non-interpretable beat sequences. Although this may exclude some low-quality or highly pathological intervals, it reduces systematic noise from peak-detection errors in downstream analysis.

From the refined R-peaks, we computed beat-to-beat RR intervals in seconds and applied a 3-beat median filter to suppress isolated outliers caused by occasional detection errors [23]. The resulting RR interval sequences capture heart rate oscillations associated with apnea [24]. We also extracted the ECG amplitude at each refined R-peak to form a beat-level amplitude series. This feature reflects respiration-related modulation of QRS amplitude and supports ECG-derived respiration analysis [25]. The amplitude series was aligned with the RR interval sequence for each window.

Additional quality checks removed windows with mean or instantaneous heart rates outside 20 to 300 bpm and windows with erratic RR interval behavior consistent with missed beats, motion artifacts, or sensor issues. Only segments with stable, interpretable cardiac dynamics were retained for downstream processing [26]. Each window inherited the apnea or normal label of its central minute, which avoids ambiguity in overlapping windows and matches common apnea-detection protocols [20]. Labels were stored as binary values (0 normal, 1 apnea), and subject identifiers were retained to enable subject-disjoint training and evaluation. Because RR interval and amplitude series are beat-timed and irregularly sampled, we resampled both onto a uniform 3 Hz grid over each 300-second window (900 time points). We applied per-window min–max normalization to each channel before linear interpolation, producing fixed-length sequences while preserving low-frequency HRV and respiration-linked amplitude trends [27]. The same procedure was applied to training and test data. The resampled RR interval and R-peak amplitude sequences were stacked to form a two-channel input tensor of shape (900, 2) per window. This representation jointly encodes heart rate variability and ECG-derived respiration, two complementary markers associated with sleep apnea [25] [24]. All preprocessing was applied consistently across splits to yield standardized inputs for model training and evaluation. See Fig. 1.

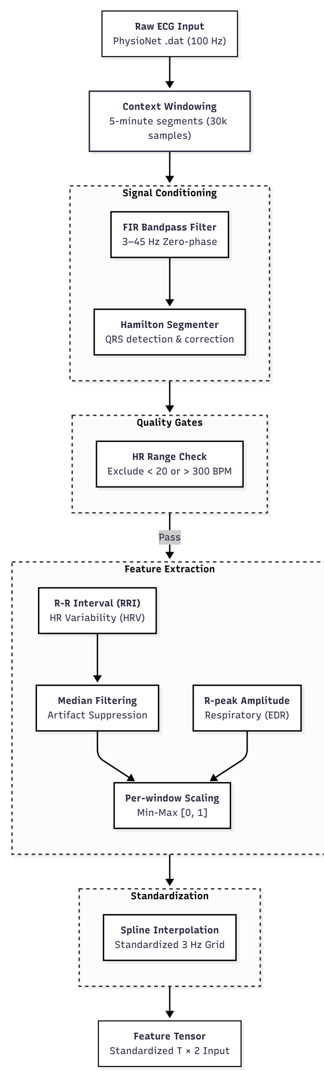

Fig. 1. Preprocessing pipeline for ECG-based apnea classification.

C. Model Architecture

We used a hybrid architecture consisting of a 1D convolutional neural network (CNN) followed by a transformer encoder block. The CNN contains three convolutional layers (64, 128, and 128 filters; kernel size 7; stride 1), each followed by a ReLU activation and max-pooling with pool size 4, which progressively compresses the temporal dimension while expanding the feature representation. A dropout rate of 0.5 is applied after the final CNN layer to reduce overfitting. The CNN takes an input tensor of shape `(num_samples, sequence_length, num_features)` and operates jointly across feature channels.

The output of the CNN is passed to a transformer encoder block. The encoder includes multi-head self-attention (2 heads, key dimension 32) with learnable sinusoidal positional encoding added to the normalized input. Residual connections are used around both the attention and feed-forward sublayers, each followed by layer normalization. The feed-forward network consists of two dense layers (128 units each). A dropout rate of 0.5 is applied to the transformer output. The main difference from a traditional transformer encoder is that the input is normalized prior to the self-attention layer.

The final transformer output is flattened and fed into a fully connected layer (128 units, ReLU), followed by a second fully connected layer with a softmax output for binary classification.

Each minute in the recording (except the first and last two) are classified by the model. These classifications as well as per-minute features (RR-interval, RR-amplitude). See Fig. 2.

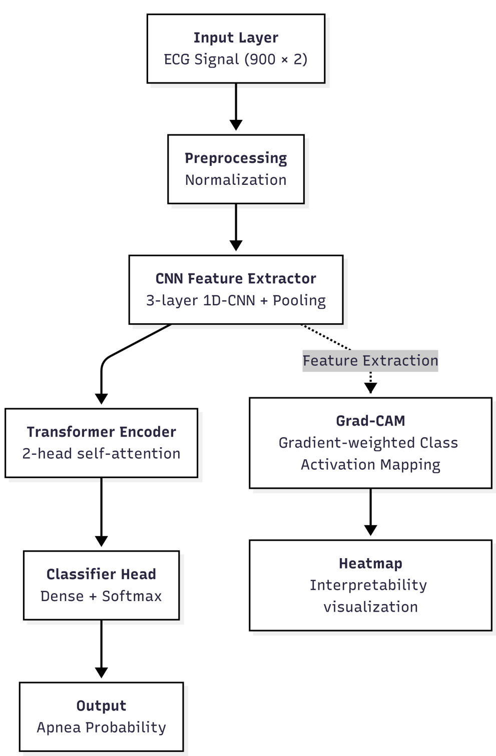

Fig. 2. CNN–Transformer architecture for minute-level apnea classification.

D. System Overview

As shown in Fig. 3, the triage system follows a three-tier architecture separating user interaction, orchestration, and model execution. This design isolates latency-sensitive risk stratification from optional interpretability and reporting modules, ensuring responsive triage while preserving transparency. A centralized orchestrator manages preprocessing, CNN-Transformer inference, and on-demand Grad-CAM generation, with outputs persisted for traceable review. Optional integration with an external language model produces structured summaries without coupling natural-language synthesis to the core inference path. By decoupling prediction, explanation, and reporting, the system supports scalable, human-in-the-loop deployment. See Fig. 3.

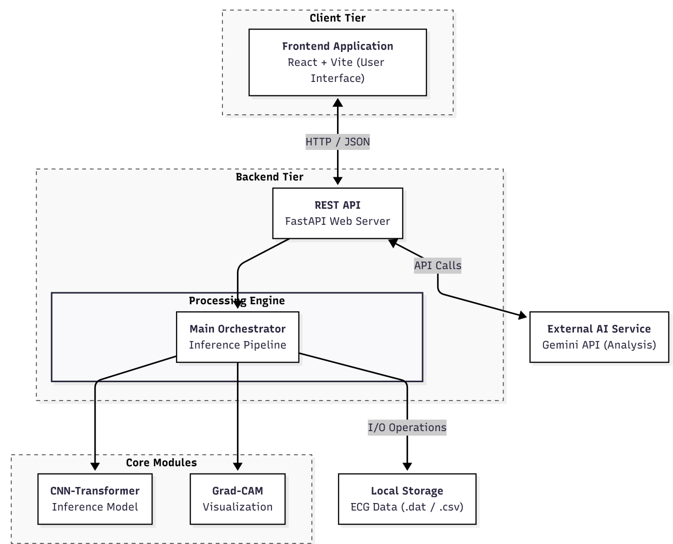

Fig. 3. System architecture for ECG-based apnea triage.

III. Results

A. Evaluation Setup

We used the PhysioNet Apnea-ECG dataset and retained the original Computing in Cardiology (CinC) competition split, using the `a/b/c` recordings (`a01–a20`, `b01–b05`, `c01–c10`) for model development and the `x` recordings (`x01–x35`) as a held-out test set [18]. Within the development set, we performed 5-fold `GroupKFold` cross-validation at the record level (28 records for training and 7 for validation per fold) to select the final model, using validation loss as the selection criterion.

B. Performance Metrics

Using the selected model, performance on the held-out evaluation set reached 89.21% accuracy, 84.24% sensitivity, 92.30% specificity, 85.68% F1-score, and an AU-ROC of 95.60%. The resulting confusion matrix (TP=5,467; TN=9,650; FP=805; FN=1,023) is shown in Fig. 4, and the ROC curve is shown in Fig. 5.

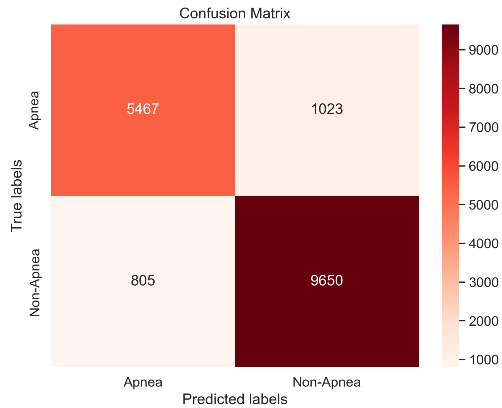

Fig. 4. Confusion matrix on the evaluation set.

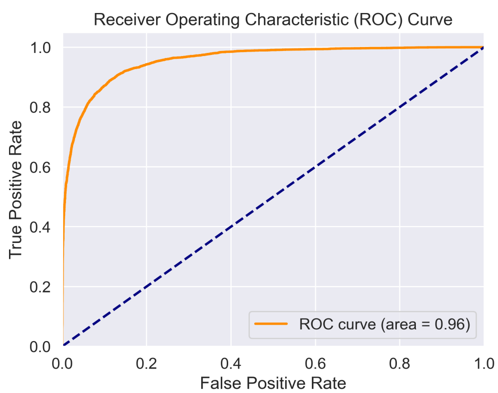

Fig. 5. ROC curve on the evaluation set.

C. Training Dynamics

Training and validation loss and accuracy over 20 epochs are shown in Fig. 6. Performance improved rapidly in the first few epochs, followed by a plateau in validation metrics while training continued to improve, suggesting mild overfitting in later epochs. The checkpoint used for downstream triage was selected at the minimum validation loss (Fold 5, epoch 11).

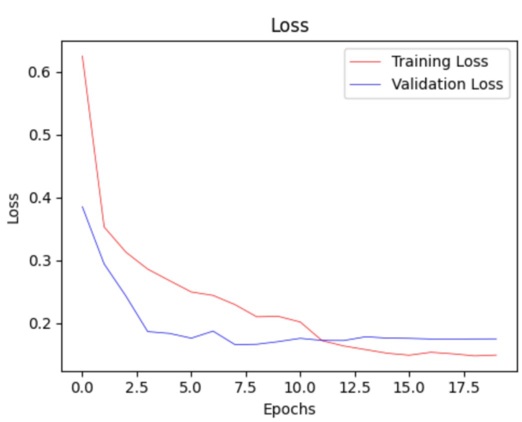

Fig. 6. Loss on train and evaluation set.

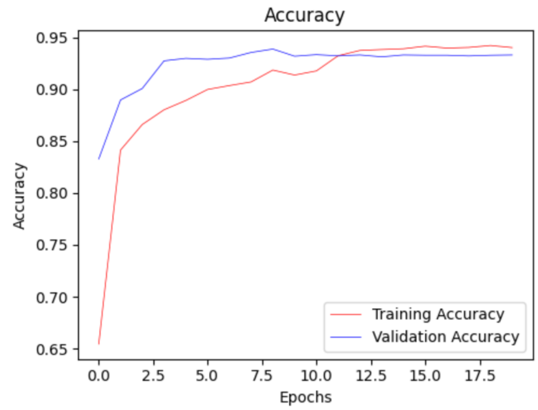

Fig. 7. Accuracy on train and evaluation set.

D. Qualitative Explainability Outputs

To support clinician-facing triage review, we used Grad-CAM as a qualitative interpretability tool rather than a performance metric. Grad-CAM heatmaps were computed from the final convolutional feature layer and overlaid on the corresponding ECG trace, where warmer regions indicate time intervals that contributed more strongly to the predicted class. Fig. 8 shows representative examples for a true positive (apnea) segment and a true negative (non-apnea) segment (Record [ID], minute [m]; predicted [class], $P(\text{apnea}) = [p]$; ground truth [label]), illustrating that the model's decision is driven by localized sub-intervals of the signal rather than uniform weighting across the full segment.

Within the triage workflow, clinicians can generate Grad-CAM overlays for any selected minute under review, accompanied by a concise per-minute markdown report summarizing the prediction, confidence, and a plausibility-oriented interpretation of the highlighted regions. These explanations are post hoc and intended for human inspection; they do not modify model predictions or contribute to the reported classification metrics.

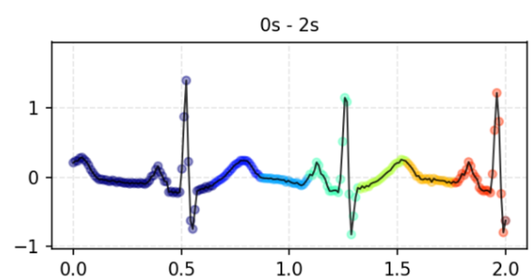

Fig. 8. Grad-CAM visualization for a representative ECG segment (0–2 s).

IV. Discussion

The selected CNN-Transformer achieved strong discrimination performance on the Apnea-ECG evaluation set (AUROC 0.956), with balanced operating-point performance (accuracy 0.892, sensitivity 0.842, specificity 0.923). A key reason AUROC is emphasized in this setting is that it summarizes ranking quality across thresholds, which is especially relevant for triage workflows where the threshold can be tuned to match clinical priorities (maximizing sensitivity for "don't miss" screening vs. increasing specificity to reduce unnecessary review). In practical triage, minimizing false positives matters because excessive flagging increases clinician workload and can contribute to alert fatigue; the observed specificity suggests a relatively low false-positive rate at the default decision rule, and the high AUROC indicates there is room to shift the threshold toward fewer false alarms while preserving meaningful sensitivity. Our evaluation follows the standard released/withheld protocol of the Apnea-ECG database (`a/b/c` for development; `x` for testing), enabling direct comparison with prior studies that adopt the same split. From a systems diagnostics perspective, these results support the model's role as an early triage tool rather than a standalone diagnostic system. It can prioritize high-risk minutes for review, enabling more efficient downstream assessment and definitive diagnosis when needed. This framing is important because a model can be clinically useful despite being imperfect, provided it is deployed as decision support with transparent uncertainty and a controllable operating point.

Failure cases are expected to cluster around signal-quality and preprocessing sensitivities. In particular, noisy segments, motion artifacts, and conditions that degrade R-peak detection or distort RR-interval dynamics can lead to misclassification, since the pipeline relies on consistent beat detection and derived features. More broadly, minute-level labels can be inherently ambiguous near transitions or in borderline events, which can produce errors even for otherwise well-performing classifiers.

To improve interpretability and support human review, Grad-CAM visualizations provide qualitative evidence that predictions are driven by localized sub-intervals rather than treating each segment uniformly. In the triage workflow, the explainability agent enables physicians to generate Grad-CAM overlays for any selected minute, pairing model confidence with a concise plausibility-oriented summary. These explanations are intended to increase transparency and review efficiency; however, they remain post-hoc and do not establish causality, and should be interpreted as supporting context rather than physiological proof. There are several limitations. First, the dataset size is modest at the recording level (70 total, with 35 held out), and performance may not fully reflect real-world variability across devices, populations, and comorbidities. Second, the class balance ($\approx$38% positive at the sample level) is not extreme but still relevant, and threshold choice affects how false positives vs. false negatives trade off in practice. Third, generalization beyond Apnea-ECG remains untested; external validation is required before clinical conclusions can be drawn.

Future work should focus on (i) improving robustness to noisy ECG and beat-detection failures, (ii) tuning and calibrating probability outputs so thresholds can be set to explicit clinical targets (e.g., "keep false positives below X while maintaining sensitivity"), and (iii) expanding explainability from single-case visuals into lightweight summaries that can be reviewed quickly at scale. In addition, a small human usefulness study could assess whether the agent's reports reduce review time and increase trust without encouraging overreliance.

V. Conclusion

Neural Sleep Apnea Triage evaluates whether single-lead overnight ECG can support scalable sleep apnea screening when polysomnography access is limited. Using the PhysioNet Apnea-ECG dataset, ECG recordings were converted into 5-minute context windows and represented as paired R-R interval and R-peak amplitude sequences. A hybrid CNN-Transformer model then produced minute-level apnea probabilities intended for triage and prioritization, not diagnosis.

On the held-out evaluation set, the system achieved AUROC 0.956 with 84.24% sensitivity and 89.21% accuracy, indicating strong discrimination and suitability for threshold-tuned, high-sensitivity screening workflows.

These results support ECG as a practical first-line screening layer that can help route higher-risk patients toward confirmatory polysomnography, reducing diagnostic bottlenecks without replacing PSG as the definitive test. Next steps include external validation across devices and populations, improved robustness to noisy signals and beat-detection failures, probability calibration for clinically defined operating points, and prospective evaluation of workflow impact and interpretability usefulness.

References

[1] G. Iannella, A. Pace, M. G. Bellizzi, G. Magliulo, A. Greco, A. De Virgilio, E. Croce, F. M. Gioacchini, M. Re, A. Costantino, M. Casale, A. Moffa, J. R. Lechien, S. Cocuzza, C. Vicini, A. Caranti, R. Marchese Aragona, M. Lentini, and A. Maniaci, "The global burden of obstructive sleep apnea," *Diagnostics*, vol. 15, no. 9, p. 1088, Apr. 2025. [Online]. Available: http://dx.doi.org/10.3390/diagnostics15091088

[2] Y. Yeghiazarians, H. Jneid, J. R. Tietjens, S. Redline, D. L. Brown, N. El-Sherif, R. Mehra, B. Bozkurt, C. E. Ndumele, and V. K. Somers, "Obstructive sleep apnea and cardiovascular disease: A scientific statement from the american heart association," *Circulation*, vol. 144, no. 3, Jul. 2021. [Online]. Available: http://dx.doi.org/10.1161/CIR.0000000000000988

[3] D. Pacheco, "How much does a sleep study cost?" https://www.sleepfoundation.org/sleep-studies/how-much-does-a-sleep-study-cost, 2025, accessed: 2025-12-04.

[4] W. W. Flemons, N. J. Douglas, S. T. Kuna, D. O. Rodenstein, and J. Wheatley, "Access to diagnosis and treatment of patients with suspected sleep apnea," *American Journal of Respiratory and Critical Care Medicine*, vol. 169, 2004. [Online]. Available: https://www.atsjournals.org/doi/pdf/10.1164/rccm.200308-1124pp

[5] T. Penzel and M. Salanitro, "Emerging challenges in the transition from manual to automated sleep scoring," *SLEEPJ*, vol. 48, no. 10, Jul. 2025. [Online]. Available: http://dx.doi.org/10.1093/sleep/zsaf202

[6] B. Uznańska, E. Trzos, T. Rechciński, J. D. Kasprzak, and M. Kurpesa, "Repeatability of sleep apnea detection in 48-hour holter ecg monitoring," *Annals of Noninvasive Electrocardiology*, vol. 15, no. 3, p. 218–222, Jul. 2010. [Online]. Available: http://dx.doi.org/10.1111/j.1542-474X.2010.00367.x

[7] M. Kolhar, M. M. Alfridan, and R. A. Siraj, "Ai-driven detection of obstructive sleep apnea using dual-branch cnn and machine learning models," *Biomedicines*, vol. 13, no. 5, p. 1090, Apr. 2025. [Online]. Available: http://dx.doi.org/10.3390/biomedicines13051090

[8] M. E. Kilic, M. E. Arayici, O. E. Turan, Y. R. Yilancioglu, E. E. Ozcan, and M. B. Yilmaz, "Diagnostic accuracy of machine learning algorithms in electrocardiogram-based sleep apnea detection: A systematic review and meta-analysis," *Sleep Medicine Reviews*, vol. 81, p. 102097, Jun. 2025. [Online]. Available: http://dx.doi.org/10.1016/j.smrv.2025.102097

[9] Y. Nygate, M. Sprague, S. Rusk, C. Fernandez, and N. Watson, "0680 clinical validation of ECG-based obstructive sleep apnea screening using machine learning," *SLEEP*, vol. 48, no. Supplement 1, pp. A296–A296, May 2025. [Online]. Available: http://dx.doi.org/10.1093/sleep/zsaf090.0680

[10] D. T. Pham and R. Mouček, "Efficient sleep apnea detection using single-lead ecg: A cnn-transformer-lstm approach," *Computers in Biology and Medicine*, vol. 196, p. 110655, Sep. 2025. [Online]. Available: http://dx.doi.org/10.1016/j.compbiomed.2025.110655

[11] P. Biswas and M. A. Yousuf, "Leveraging transformer models for accurate detection of obstructive sleep apnea from single-lead ecg signals," in *Proceedings of the 3rd International Conference on Computing Advancements*, ser. ICCA 2024. ACM, Oct. 2024, p. 556–563. [Online]. Available: http://dx.doi.org/10.1145/3723178.3723252

[12] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, and J. H. Peter, "The apnea-ecg database," *Computers in Cardiology*, 2000. [Online]. Available: http://georgebmoody.com/publications/apnea-ecg-cinc-2000.pdf

[13] T. Paul, O. Hassan, C. S. McCrae, S. K. Islam, and A. S. M. Mosa, "An explainable fusion of ecg and spo2-based models for real-time sleep apnea detection," *Bioengineering*, vol. 12, no. 4, p. 382, Apr. 2025. [Online]. Available: http://dx.doi.org/10.3390/bioengineering12040382

[14] R. Haidar, S. McCloskey, I. Koprinska, and B. Jeffries, "Convolutional neural networks on multiple respiratory channels to detect hypopnea and obstructive apnea events," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–7.

[15] A. Rizal, F. D. A. A. Siregar, and H. T. Fauzi, "Obstructive sleep apnea (osa) classification based on heart rate variability (hrv) on electrocardiogram (ecg) signal using support vector machine (svm)," *Traitement du Signal*, vol. 39, no. 2, p. 469–474, Apr. 2022. [Online]. Available: http://dx.doi.org/10.18280/ts.390208

[16] H. T. Chaw, S. Kamolphiwong, and K. Wongsritrang, "Sleep apnea detection using deep learning," *Tehnički glasnik*, vol. 13, no. 4, p. 261–266, Dec. 2019. [Online]. Available: http://dx.doi.org/10.31803/tg-20191104191722

[17] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, and J. H. Peter, "Apnea-ecg database," 2000. [Online]. Available: https://physionet.org/content/apnea-ecg/

[18] G. Moody and R. Mark, "Detecting and quantifying apnea based on the ecg: The physionet/computing in cardiology challenge 2000," https://physionet.org/content/challenge-2000/1.0.0/, 2003, accessed: 2025-12-04.

[19] G. B. Moody and R. G. Mark, "Computers in cardiology challenge 2000: The apnea-ECG database," https://moody-challenge.physionet.org/2000/, 2000, accessed: 2025-12-04.

[20] C.-K. Peng, "Apnea detection from the ecg," https://physionet.org/content/apdet/1.0.0/, 2002, accessed: 2025-12-04.

[21] Z. Wang, Q. Zhang, K. Lan, Z. Yang, X. Gao, A. Wu, Y. Xin, and Z. Zhang, "Enhancing instantaneous oxygen uptake estimation by non-linear model using cardio-pulmonary physiological and motion signals," *Frontiers in Physiology*, vol. 13, Aug. 2022. [Online]. Available: http://dx.doi.org/10.3389/fphys.2022.897412

[22] P. S. Hamilton, "Open source ECG analysis," in *Computers in Cardiology 2002*. IEEE, 2002, pp. 101–104. [Online]. Available: https://www.cinc.org/old/Proceedings/2002/pdf/101.pdf

[23] Kubios Oy, "Preprocessing of HRV data," https://www.kubios.com/blog/preprocessing-of-hrv-data/, accessed: 2025-12-04.

[24] T. Penzel, J. McNames, P. de Chazal, B. Raymond, A. Murray, and G. Moody, "Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings," *Medical and Biological Engineering and Computing*, vol. 40, no. 4, p. 402–407, Jul. 2002. [Online]. Available: http://dx.doi.org/10.1007/BF02345072

[25] G. B. Moody, R. G. Mark, M. A. Bump, J. S. Weinstein, A. D. Berman, J. E. Mietus, and A. L. Goldberger, "Clinical validation of the ECG-derived respiration (EDR) technique," in *Computers in Cardiology 1986*. Washington, DC: IEEE Computer Society Press, 1986, pp. 507–510. [Online]. Available: https://archive.physionet.org/physiotools/edr/cic86/

[26] K. J. Kemper, C. Hamilton, and M. Atkinson, "Heart rate variability: Impact of differences in outlier identification and management strategies on common measures in three clinical populations," *Pediatric Research*, vol. 62, no. 3, p. 337–342, Sep. 2007. [Online]. Available: http://dx.doi.org/10.1203/PDR.0b013e318123fbcc

[27] F. Gasparini, A. Grossi, M. Giltri, and S. Bandini, "Personalized ppg normalization based on subject heartbeat in resting state condition," *Signals*, vol. 3, no. 2, p. 249–265, Apr. 2022. [Online]. Available: http://dx.doi.org/10.3390/signals3020016