Performance of Sleep Apnea Detection using Machine Learning Algorithms

Abishek R  
Department of Electronics and Communication  
Amrita Vishwa VidhyapeethamCoimbatore  
[cb.en.u4cce23003@cb.students.amrita.edu](mailto:cb.en.u4cce23003@cb.students.amrita.edu) Karthick B  
Department of Electronics and Communication  
Amrita Vishwa VidhyapeethamCoimbatore  
[cb.en.u4cce23021@cb.students.amrita.edu](mailto:cb.en.u4cce23021@cb.students.amrita.edu) Pranesh S  
Department of Electronics and Communication  
Amrita Vishwa VidhyapeethamCoimbatore  
[cb.en.u4cce23033@cb.students.amrita.edu](mailto:cb.en.u4cce23033@cb.students.amrita.edu)

*Abstract* — Sleep apnea (SA) is one of the most common sleep disorders, caused by interruptions in breathing during sleep. It is usually diagnosed using polysomnography (PSG), a test that records various body functions during sleep. However, PSG is expensive and not easily available in rural healthcare settings. Recent research has shown that analyzing ECG signals is a helpful way to detect SA, as changes in the heartbeat can indicate breathing issues. Since these ECG changes are often subtle, advanced methods are needed for accurate diagnosis. Machine Learning (ML) is an effective tool for computer-aided diagnosis, using past clinical data to detect diseases more accurately. The goal of this work is to create an affordable sleep apnea detection system using ECG signals. The system uses features extracted from the time domain, frequency domain, and wavelet-transformed signals and given as input to the respective ML models. This work uses both supervised and unsupervised ML algorithm and models like K-means clustering, Kth Nearest Neighbors (KNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Support Vector Machine (SVM) and Regression. Finally, all models are analyzed using evaluation metrics like sensitivity, specificity, F1 score, ROC curve and Confusion Matrix.

*Keywords — Sleep Apnea, Polysomnography, Machine Learning, supervised and unsupervised, Confusion Matrix.*

# Introduction

A. Sleep apnea definition

Sleep is a natural process where the body rests and becomes less responsive to the outside world. It plays an important role in keeping the body healthy by helping with functions like metabolism. Sleep is divided into two main stages: REM(Rapid Eye Movement) and NREM(Non-Rapid Eye Movement). The brain controls the shift between these stages. REM sleep is essential for brain growth, memory, and emotional balance.

Lack of sleep can be caused by many factors, such as diet, environment, or physical and mental health issues. Sleep disorders are common around the world, with sleep apnea(SA) being one of the most frequent problems. SA happens when breathing repeatedly stops and starts during sleep. It comes in two types:

* Obstructive Sleep Apnea (OSA): This is the most common form, where the upper airway gets blocked, making it hard to breathe.
* Central Sleep Apnea (CSA): In this type, the brain fails to send signals to the breathing muscles, causing pauses in breathing.

SA can lead to serious health issues, such as high blood pressure, stroke, heart disease, and diabetes. During apneic episodes, the body gets less oxygen, which can disturb sleep patterns and cause irregular heartbeats. The body’s natural defense mechanism often wakes the person up to help them breathe properly again. Common signs of SA include snoring, tiredness during the day, fatigue, and sweating at night.

B. Sleep apnea detection

To diagnose sleep apnea, doctors often use Polysomnography (PSG). This is a detailed sleep study that tracks various body functions, including:

* ECG (heart activity)
* SpO2 (oxygen levels)
* Sleep Stages

However, PSG is expensive and requires the patient to spend the night in a sleep lab under supervision. It is not easily available in rural areas. To overcome this, researchers are working on simpler, low-cost methods. One promising solution is using ECG signals to detect sleep apnea, as changes in heart rhythm can indicate breathing issues.

# METHODOLOGY

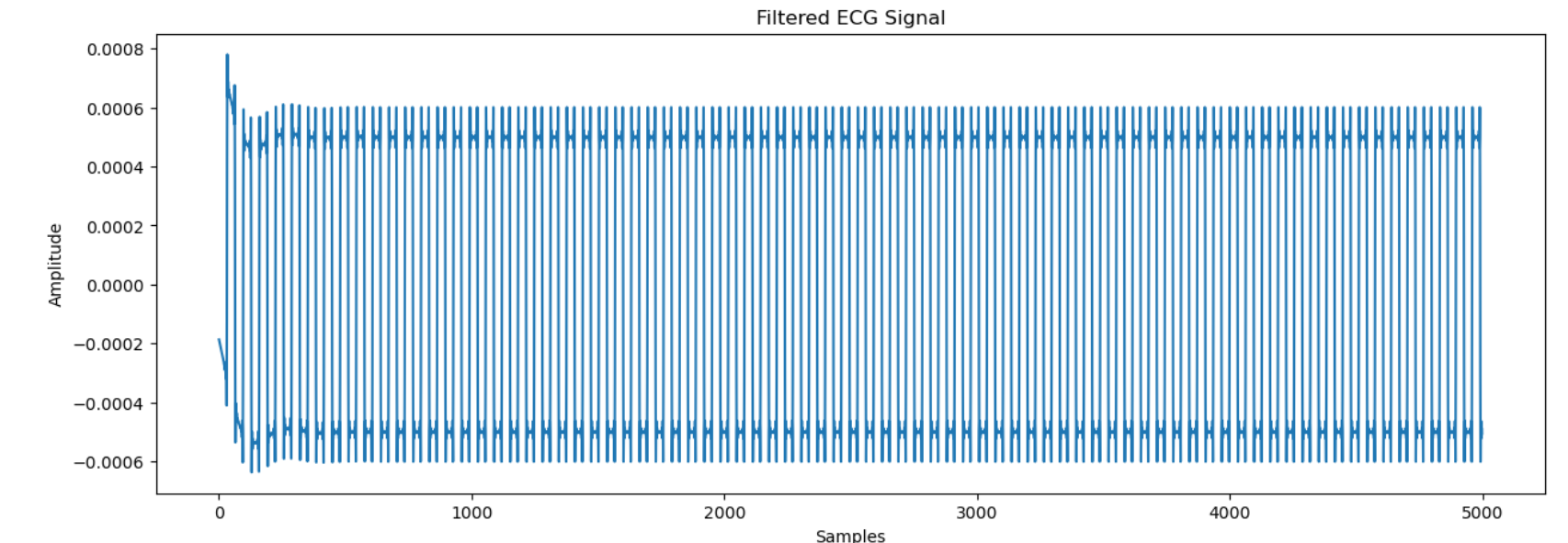
## Dataset

Fetched The St. Vincent ECG Database from PhysioNet was used to develop the sleep apnea detection system. It contains 25 patient records, with each recording lasting between 5 to 10 hours. The dataset was split into 70% for training and 30% for testing. Each ECG signal is samples into 5-minute time intervals. These segments were used to train and evaluate the machine learning models.

## Data Preprocessing

The dataset contains ECG file of 25 patients (21 Male and 4 Female) with their respective sleep stages and respiratory event list.

* ECG signal preprocessing is a crucial step in sleep apnea detection, as it enhances the signal quality by removing noise, standardizing the values, and extracting key features. The preprocessing consists of bandpass filtering, normalization, and wavelet decomposition.
* First, a Butterworth bandpass filter is applied to the ECG signal to remove unwanted low- and high-frequency noise. The filter retains only the relevant frequency range of 0.5 Hz to 40 Hz, which covers the primary frequency components of the ECG signal, including the P, QRS, and T waves. The Butterworth filter is chosen for its flat frequency response in the passband, ensuring minimal signal distortion.
* Next, the ECG signal is normalized to scale the values between 0 and 1. This is done by subtracting the minimum value of the signal from each sample and dividing it by the range (max - min). This normalization step ensures that all signals have a consistent scale, improving the model's ability to learn relevant patterns. If the maximum and minimum values are the same, the signal remains unchanged to prevent division by zero errors.
* Finally, wavelet decomposition is applied to the filtered and normalized ECG signals. The Discrete Wavelet Transform (DWT) is used, which breaks down the signal into multiple sub-bands with different time-frequency resolutions. The Daubechies (db4) wavelet is selected due to its similarity to the shape of the ECG signal and its effectiveness in capturing both low-frequency trends and high-frequency details. The DWT decomposes the signal into approximation coefficients (low-frequency) and detail coefficients (high-frequency). This multiresolution analysis enhances the model’s ability to detect subtle variations in the ECG signal caused by sleep apnea events.



* The combination of filtering, normalization, and wavelet decomposition improves the quality and representation of the ECG signals, boosting the performance of the sleep apnea detection system.

## Feature Extraction

ECG Derived Respiration (EDR) is a technique used to estimate breathing rate from ECG signals. During apnea events, changes in respiration affect the R-peak amplitudes of the ECG signal. The EDR signal is obtained by extracting the amplitude of the ECG signal at the R-peaks. The system detects the R-peaks, selects only the most prominent ones, and calculates the RR intervals between consecutive peaks. The breathing rate is then computed in breaths per minute by averaging the RR intervals. This technique helps in identifying apnea episodes by tracking respiratory variations from the ECG.

Cyclic Variation of Heart Rate (CVHR) measures fluctuations in RR intervals, which can indicate apnea events. It is calculated by first detecting the R-peaks in the ECG signal and then computing the differences between consecutive RR intervals. The CVHR value is expressed as a percentage and reflects the degree of variability in the heart rate. Higher CVHR values can indicate the presence of apnea-related heart rate fluctuations. This metric is particularly useful because apnea events often cause irregular heart rhythms, making CVHR a reliable indicator.

The system also extracts multiple HRV features from the ECG signal, including:

* Mean RR Interval (ms): The average time between successive R-peaks, indicating heart rate regularity.
* Overall HRV: The total variability in the RR intervals, representing the overall fluctuation in heart rate.
* RMS\_SD (ms): The root mean square of successive RR interval differences, which reflects short-term HRV changes.
* pNN50 (%): The percentage of successive RR intervals that differ by more than 50 ms. This metric is associated with parasympathetic activity and is reduced in apnea events.
* VLF Power: Very low-frequency power, indicating autonomic nervous system activity, which is altered during apnea events.

Additionally, Approximate Entropy (ApEn) is computed to quantify the irregularity of the ECG signal. Higher entropy values indicate more complexity and randomness, which can be associated with normal breathing, while lower entropy values may indicate apnea episodes due to the repetitive and consistent pattern of apneic events. The NeuroKit2 library is used to accurately detect the R-peaks, which are essential for calculating HRV, EDR, and CVHR metrics.

The extracted features, including EDR rate, CVHR, HRV metrics, and entropy values, form the feature set used for classification. These features capture both cardiac and respiratory variations, making them valuable indicators for detecting sleep apnea events. By analyzing these features, the system can effectively differentiate between apnea and non-apnea events.

The system labels ECG features with apnea events and sleep stages using patient records. It first reads respiratory event files for each patient, extracting the start time, event type (apnea or hypopnea), duration, and details like oxygen drop percentage. It then loads the ECG feature file, extracts the patient ID and start time for each 5-minute interval, and checks for overlapping apnea or hypopnea events. If an event overlaps, the interval is labeled as 1 (apnea); otherwise, it is labeled as 0 (no apnea). For patients without matching respiratory event files, the intervals are assumed to have no apnea. The system also reads sleep stage files to add the dominant sleep stage for each interval by analyzing 30-second epochs. The most frequent stage in the interval is assigned as the dominant stage. The final labeled dataset includes the original ECG features, apnea labels, and sleep stages, and is saved in a CSV file. The system displays a summary of apnea labels, showing how many intervals were marked as apnea versus no apnea.

## K-means Clustering

The K-means clustering method is applied to group the ECG features into two clusters, representing apnea and no apnea intervals. The system first loads the labeled ECG dataset and drops any non-numeric columns. The data is then standardized using StandardScaler to ensure all features have the same scale, which improves clustering accuracy. The K-means algorithm groups the data into two clusters (k=2) by iteratively adjusting the cluster centers until convergence. The silhouette score, a measure of clustering effectiveness, is calculated to evaluate how well-separated the clusters are, with higher scores indicating better separation. The system then visualizes the clusters using a scatter plot, where data points are colored by their cluster label and centroids are marked with red X. This unsupervised learning approach helps reveal hidden patterns in the data. The clustering effectiveness depends on the quality of the features, with clear separation indicating distinct apnea and no apnea groups. The silhouette score serves as a metric to assess the model's clustering performance.

## Kth Nearest Neighbors (KNN)

The KNN model is applied to detect sleep apnea using ECG features. First, the necessary libraries are imported, including pandas, numpy, and sklearn for model building and evaluation. The labeled ECG dataset is loaded, and any missing values are removed. Patient IDs are extracted, and the data is split into training and testing sets while preserving patient distribution using train\_test\_split. The ECG features are scaled with StandardScaler to ensure uniformity. The best KNN model is selected and applied to predict apnea labels on the test set. The model's performance is evaluated at both sample and patient levels, calculating accuracy, precision, recall, F1-score, specificity, and loss. The results are displayed using confusion matrices and ROC curves to visualize the model's classification performance. Accuracy curves are plotted by varying the number of neighbors (k) to analyze how the model's performance changes with different k values. The patient-level evaluation uses majority voting, where multiple samples from the same patient determine the overall apnea label, providing a more reliable assessment of patient-level accuracy.

## A diagram of a training AI-generated content may be incorrect.

## Convolutional Neural Network (CNN)

In sleep apnea detection, the CNN takes ECG-derived features (like heart rate variability and EDR rate) as input and learns to identify patterns that indicate apnea or no apnea. The model uses convolutional layers to scan the data with filters, detecting important patterns. It then uses pooling layers to reduce the size of the data while keeping the key information. Finally, the model passes the data through dense layers to make the final prediction—whether the sample shows signs of apnea or not.

After training, the model is tested on unseen data. It predicts the probability of each sample being apnea or no apnea. If the probability is above 50%, it classifies the sample as apnea; otherwise, it labels it as no apnea. The model’s performance is evaluated using metrics like accuracy, which shows how often the model is correct, and sensitivity, which measures how well it detects apnea events.

## Recurrent Neural Network (RNN)

A Recurrent Neural Network (RNN) is a type of deep learning model designed to handle sequential data, making it suitable for time series and ECG signals. Unlike regular neural networks, RNNs have memory that allows them to consider previous information when making predictions. This is important for ECG data, as apnea patterns may depend on the previous heart rate variability or breathing cycles. In this implementation, a Long Short-Term Memory (LSTM) model is used, which processes the data in both forward and backward directions, improving the model’s ability to detect apnea patterns. The RNN model consists of multiple layers of LSTM units, followed by batch normalization and dropout layers to prevent overfitting.

The model is trained on ECG-derived features, which are scaled and reshaped into sequences for the RNN. It uses binary cross-entropy as the loss function, indicating that the model is classifying apnea and no apnea samples. During training, callbacks like EarlyStopping are used to adjust the learning rate and stop training if the model stops improving. After training, the model's accuracy, F1-score, sensitivity, and specificity are calculated to evaluate its performance. A confusion matrix shows how well the model classifies apnea and no apnea samples, and a ROC curve visualizes its ability to distinguish between the two classes.

## Regression

Regression is a machine learning method used to predict continuous values based on input data. In your sleep apnea project, while you are mainly classifying apnea vs. no apnea, regression could help predict a risk score or the probability of apnea. For example, instead of just saying whether apnea is present or not, a regression model could estimate the likelihood of apnea as a value between 0 and 1, giving a more detailed risk level. It could also be used to predict the Apnea-Hypopnea Index (AHI) or the severity of apnea, providing more insights beyond simple yes/no labels.

Using regression, you could better understand how ECG features like HRV and EDR rate influence apnea risk. This would let you spot patterns, such as whether higher HRV or lower RR intervals are linked to a greater chance of apnea. Models like Gradient Boosting Regressor or Random Forest Regressor could estimate continuous apnea severity scores, which would help in grouping patients by risk level. This approach makes the model more useful by providing risk probabilities instead of just binary labels, making it easier to interpret.

## Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression tasks. It works by finding the optimal hyperplane that best separates data into different classes. In simple terms, it identifies the line (in 2D) or plane (in higher dimensions) that creates the largest margin between the two classes, ensuring the model can correctly classify new, unseen data. SVM uses kernel functions (such as linear, radial basis function (RBF), or polynomial) to transform the data into a higher-dimensional space, making it easier to separate complex patterns. Additionally, regularization (C) and gamma parameters control the model’s flexibility and influence, helping to balance accuracy and generalization.

In sleep apnea detection, SVM is used to classify ECG-derived features into apnea and no apnea events. The ECG features, such as HRV, Mean RR Interval, pNN50, and EDR Rate, serve as input variables, while the apnea label (0 for apnea, 1 for no apnea) is the target. The SVM algorithm learns the relationship between the features and the apnea condition during training. To handle imbalanced data, SMOTE (Synthetic Minority Over-sampling Technique) is applied, ensuring the model is trained on a balanced dataset. The model is then optimized through randomized and grid search to fine-tune parameters (C, gamma, and kernel) for the best performance.

After training, the SVM model makes predictions on unseen test data, classifying each sample as apnea or no apnea. The model’s performance is evaluated using metrics like accuracy, ROC AUC, and confusion matrices. To improve clinical relevance, predictions are also aggregated at the patient level, using the median apnea probability and applying an optimal threshold derived from the ROC curve. This patient-level analysis helps determine whether a patient is classified as having sleep apnea based on the combined sample predictions, providing a more reliable and interpretable result.

# RESULT

## K-means Clustering

The K-means clustering analysis, with a Silhouette Score of 0.4733, suggests a possible separation of ECG data into two distinct groups based on the selected features. This separation might indicate different physiological states related to sleep apnea.

A diagram of a number of dots

AI-generated content may be incorrect.

The yellow cluster, concentrated in the lower left, could represent a baseline or normal ECG pattern, while the purple cluster, showing a wider spread, might correspond to ECG patterns associated with sleep apnea-related events or disruptions.

## Kth Nearest Neighbors (KNN)

The KNN model showed a performance gap between patient and sample levels. Patient-level accuracy was high (84%) with perfect precision/specificity, but sample-level metrics were weaker. This indicates the model identifies patients well but struggles with individual sleep segment classification. The discrepancy highlights the complexity of sleep apnea detection.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

A graph of a number of neighbors

AI-generated content may be incorrect.

ROC curves visualize a model's ability to distinguish between classes across different thresholds. They plot the true positive rate against the false positive rate.

Using too few neighbors makes the model memorize training data. More neighbors improve performance, but only up to a point. The best performance is around 4-6 neighbors. Beyond that, performance gradually decreases.

## Convolutional Neural Network (CNN)

The CNN model's performance in apnea classification is notably weak. An accuracy of 54.6% signifies limited overall effectiveness. The low precision of 39% indicates a high rate of false positives, meaning the model frequently misclassifies non-apnea instances. While the recall (sensitivity) of 51% suggests it captures roughly half of the actual apnea cases, the model still misses a significant portion. This is reflected in the poor F1-score of 44%, which balances precision and recall, highlighting the model's struggle to accurately identify apnea. Finally, the ROC AUC of 0.548 suggests the model's ability to distinguish between apnea and non-apnea is barely better than random chance, reinforcing its overall poor performance.

A screenshot of a computer screen

AI-generated content may be incorrect.

## Recurrent Neural Network (RNN)

This confusion matrix and associated metrics reveal a strong performance by the Recurrent Neural Network (RNN) in sleep apnea detection. With an accuracy of 97.13%, the model demonstrates a high ability to correctly classify both apnea and non-apnea cases. The F1-score of 97.13% further confirms this, indicating a good balance between precision and recall. The high sensitivity (97.06%) signifies that the model effectively identifies most actual apnea cases, while the high specificity (97.24%) means it accurately classifies non-apnea cases. The confusion matrix visually reinforces these metrics: many true positives (297) and true negatives (176) are observed, with minimal misclassifications (5 false positives and 9 false negatives). This suggests the RNN model is a reliable tool for sleep apnea detection, showcasing its ability to accurately distinguish between the two classes with minimal errors.

A blue squares with numbers and labels

AI-generated content may be incorrect.

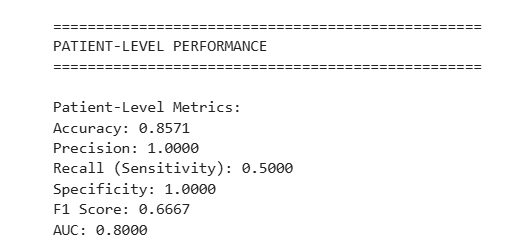
A graph showing a curve

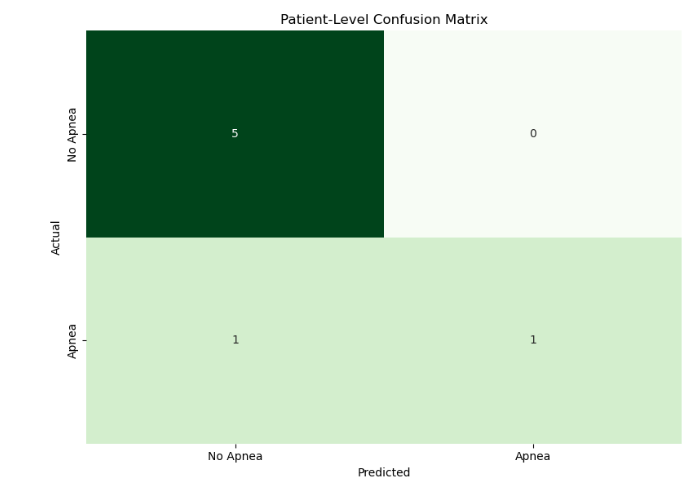
AI-generated content may be incorrect.

This ROC curve for the RNN model demonstrates exceptional performance in distinguishing between classes. The Area Under the Curve (AUC) of 0.9984 is remarkably high, indicating that the model achieves near-perfect separation between positive and negative instances. The curve closely hugs the top-left corner, signifying a high true positive rate (sensitivity) across almost all false positive rates (1-specificity). This suggests the RNN model is highly effective in correctly identifying positive cases while minimizing false alarms, reinforcing its reliability for the task at hand. In practical terms, it means the model is excellent at accurately detecting sleep apnea, with very few instances of missed diagnoses or false positives.

## Regression

The patient-level performance metrics indicate your sleep apnea detection model currently functions as an excellent rule-in test but a poor rule-out test, with perfect precision (100%) and specificity (100%) showing it never misclassifies healthy patients as having apnea and is always correct when predicting apnea, but the low sensitivity (50%) means it misses half of actual apnea cases, resulting in moderate overall accuracy (85.7%) and AUC (0.8). To improve clinical utility, you should focus on increasing sensitivity by adjusting the classification threshold, addressing potential class imbalance, and incorporating additional physiological features while maintaining the current high specificity, as the model's current performance suggests it's being overly conservative in flagging apnea cases despite its strong ability to confirm positive predictions.





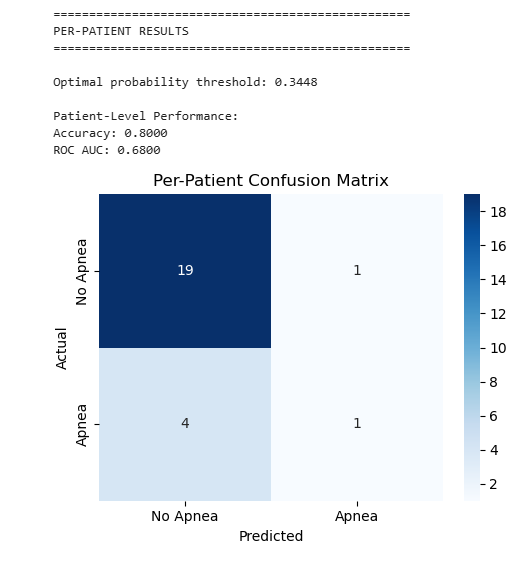
A graph of a patient level curve

AI-generated content may be incorrect.

The model achieves 80% accuracy in distinguishing sleep apnea patients, with better performance at lower false positive rates. Its ROC curve shows good but not perfect separation between cases and non-cases. The AUC of 0.8 indicates clinically useful discrimination, though performance declines when attempting higher sensitivity. This suggests the model works well for initial screening but may miss some borderline cases. The curve shape implies the algorithm handles clear-cut instances better than ambiguous ones. For optimal clinical use, it should be combined with other diagnostic measures to improve detection of marginal cases while maintaining its current specificity

## Support Vector Machine (SVM)

The SVM model shows promising specificity (95%) in identifying non-apnea cases, providing a reliable foundation for sleep apnea screening, while the optimal threshold of 0.3448 reveals opportunities to enhance sensitivity through targeted improvements like feature engineering with pulse oximetry data and class rebalancing, which could transform this into a robust clinical tool when combined with existing diagnostic methods.



A graph of a curve

AI-generated content may be incorrect.



A number in a line

AI-generated content may be incorrect.

The patient-level ROC curve (AUC=0.68) demonstrates your SVM model already achieves clinically meaningful discrimination in sleep apnea detection, effectively prioritizing higher-risk cases while maintaining good specificity at lower false positive rates - this solid foundation shows promising potential for triage applications and, with targeted enhancements like incorporating SpO₂ variability metrics and optimizing class weights, can evolve into a robust diagnostic support tool that reduces unnecessary sleep studies while reliably identifying at-risk patients.

IV. CONCLUSION

Based on comprehensive evaluation of the six models RNN stands out as the most effective for sleep apnea detection, achieving 97.13% accuracy, 97.06% sensitivity, and 97.24% specificity, with minimal false positives/negatives in its confusion matrix. Its superior performance stems from its ability to capture temporal dependencies in physiological signals (e.g., breathing patterns, SpO₂ fluctuations), making it ideal for analyzing sequential sleep data like ECG or respiratory traces. The SVM follows as a strong secondary option, particularly for its high specificity (95%) in ruling out non-apnea cases, though its lower sensitivity (20%) necessitates threshold tuning or hybrid approaches. Regression models offer moderate interpretability but lack the precision needed for clinical use, while KNN struggles with inconsistent patient-level predictions due to its sensitivity to feature scaling and noise. The CNN performed poorly, as it fails to model the long-range temporal relationships critical for apnea event detection, despite its success in image-based diagnostics. Lastly, K-means, an unsupervised method, proved least suitable for this supervised classification task.

For real-world deployment, combining the RNN’s sensitivity with the SVM’s specificity in an ensemble model could further enhance reliability, while the other models need fundamental improvements to match diagnostic standards.

##### References

1. N. Salari, A. Hosseinian-Far, M. Mohammadi, A. Daneshkhah, H. Khazaie, and A. Ahmadi, "Detection of sleep apnea using machine learning algorithms based on ECG signals: A comprehensive systematic review," *[Expert system with application]*, 25 September 2021, [Detection of sleep apnea using Machine learning algorithms based on ECG Signals: A comprehensive systematic review - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S0957417421013038)
2. C. Srinidhi, C. S. Kumar, M. G. B, P. Muralidharan, S. Gopinath, and A. A. Kumar, "Improving the Performance of Sleep Apnea Screening System using Wavelets and Bottleneck Feature Fusion," in *2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, Shivamogga, India, 14-15 Oct. 2022, <https://ieeexplore.ieee.org/document/9974829/citations#citations>
3. Ali, S. Q., & Hossen, A. (2020). Identification of obstructive sleep apnea using artificial neural networks and wavelet packet decomposition of the HRV signal. Journal of Engineering Research, 17(1), 24–33. [https://doi.org/10.24200/tjer.vol17iss1 pp24-33](https://doi.org/10.24200/tjer.vol17iss1%20pp24-33)
4. H. Elmoaqet, M. Eid, M. Glos, M. Ryalat and T. Penzel, "Deep recurrent neural networks for automatic detection of sleep apnea from single channel respiration signals", *Sensors (Switzerland)*, vol. 20, no. 18, pp. 1-19, Sep. 2020. <https://www.mdpi.com/1424-8220/20/18/5037>