HMM for Sleep Apnea Detection

Chandini Karrothu
Kent State University
Kent, United States
ckarroth@kent.edu

***Abstract*—Sleep apnea is a potentially life-threatening condition marked by frequent pauses in breathing during sleep, which can result in serious health complications such as heart disease, stroke, and impaired cognitive function. This project investigates the use of Hidden Markov Models (HMM) for detecting sleep apnea events, including apneas, hypopneas, and no apneas, by analyzing polysomnographic data such as ECG signals. The HMM model, where sleep stages are treated as hidden states and ECG signals as observed inputs, was trained on a dataset consisting of labeled sleep study annotations. The model achieved a respiration event classification accuracy of 51.92%, with notable performance discrepancies across different events, such as low precision and recall for hypopneas and apneas, and relatively better performance for detecting "no apnea" events. Sleep stage decoding accuracy was found to be 11.40%, indicating challenges in accurately classifying sleep stages. The project also highlighted limitations related to imbalanced datasets, model overfitting, and the inadequacy of the selected feature set to capture complex sleep physiology. Despite these challenges, the study demonstrates the potential of HMM-based models for automated sleep apnea detection, providing a basis for further refinement in model performance, data processing, and feature engineering for more accurate, real-time predictions in clinical settings.**

***Index Terms*—Hidden Markov Model, Sleep Apnea, Preprocessing, Hidden States, Machine Learning**

I. Introduction

Sleep apnea is a serious sleep disorder characterized by repeated interruptions in breathing during sleep. These interruptions, called apneas, can last from a few seconds to minutes, and they can occur hundreds of times throughout the night. Sleep apnea leads to increased risks of cardiovascular issues, cognitive impairment, daytime fatigue, and accidents. Early diagnosis and treatment, such as CPAP therapy, are crucial to mitigating these risks. However, traditional diagnostic methods are costly and time-consuming. Machine learning (ML) offers a promising solution by analyzing polysomnographic data, including ECG, EEG, and respiratory signals, to detect sleep apnea episodes early. By recognizing subtle patterns in these signals, ML models can provide accurate, real-time predictions, enabling timely intervention and personalized treatment. This approach reduces the risk of unforeseen health complications, improves diagnostic accuracy, and makes sleep apnea management more efficient and cost-effective. Additionally, Hidden Markov Models (HMM) are applied in this analysis, where sleep stages are treated as hidden states and ECG signals are used as observed inputs. The target values, including Hypopneas, Apneas, and No Apneas, allow the model to detect these events based on the transitions in sleep stages and heart rhythms, further enhancing detection accuracy and supporting more effective, personalized treatment strategies.

II. Data Analysis and Preparation

A. Dataset Overview

The MIT-BIH Polysomnographic Database [1] contains over 80 hours of polysomnographic recordings from patients monitored for sleep apnea and other sleep disorders. It includes multi-channel signals such as ECG, EEG, and respiratory data, annotated with sleep stages and apnea events. The dataset consists of 18 records, each containing four files: sleep/apnea annotations (.st), beat annotations (.ecg), signal data (.dat), and header information (.hea). This rich dataset is widely used for research and development of machine learning models aimed at improving the early detection and analysis of sleep apnea and related sleep disorders. For this analysis, only the ECG signals were used to detect sleep apnea events.

B. Data-preparation

To prepare the data for analysis, the sleep study annotation files are parsed to extract and categorize sleep stages and apnea events over time. The annotations are mapped to specific sleep stages, such as Awake, Stage 1, Stage 2, Stage 3, Stage 4, and REM, while apnea events like apneas and hypopneas are identified and assigned accordingly. The parsed data is processed to ensure accurate tracking of stage transitions, filling any gaps in the stage information. The ECG signal records are then processed, and missing or unknown stages are handled through post-processing. This results in a comprehensive dataset that accurately reflects sleep patterns and apnea events for further analysis.

To process the annotations for event and sleep stage analysis, the annotation file is parsed to extract events and stages for the entire duration. The data is then divided into overlapping windows, where each window corresponds to a segment of the signal. For each window, the events and stages are identified, and the most frequent event and stage are selected based on their occurrence within the window. The process ensures that each window is classified with the most representative event and sleep stage, providing a windowed dataset that captures the temporal distribution of events and stages throughout the recording.

The features were extracted from the ECG windows to capture both statistical and frequency-domain characteristics of the signal. Here's a concise explanation of each feature, including the event and sleep stage information:

*1) **mean**:* The average value of the signal, helping to detect baseline shifts in the ECG signal.

*2) **variance**:* Measures the variability of the signal, useful for detecting irregular heartbeats or arrhythmias.

*3) **skewness**:* Quantifies the asymmetry of the signal distribution, highlighting abnormalities in heart rhythm.

*4) **kurtosis**:* Indicates the "tailedness" or sharpness of the signal's distribution, helping identify extreme variations in heart activity.

*5) **shannon_entropy**:* Measures the randomness or complexity of the signal, useful for assessing the regularity of heart rhythms.

*6) **zero_crossing**:* Counts the number of times the signal crosses zero, useful for detecting rapid heart changes or anomalies.

*7) **peak_frequency**:* Identifies the dominant frequency in the signal, helping to detect characteristic heart frequencies.

*8) **harmonic_ratio**:* The ratio of harmonic components to the DC component in the FFT, revealing the rhythmic structure of the ECG.

*9) **heart_rate**:* Calculated from the number of QRS peaks, directly measuring the number of heartbeats per minute.

*10) **event**:* Describes the event associated with the window, categorized as:

- **Hypopnea**: Partial airway blockage leading to shallow breathing.
- **Apnea**: Complete airway blockage leading to cessation of breathing.
- **No Apnea**: No significant breathing abnormalities detected (e.g., leg movements, arousals).

*11) **sleep_stage**:* Provides the sleep or physiological stage associated with the window, with possible stages including:

- **W (Wake)**: Awake state, typically corresponding to a more irregular heart pattern.
- **1 (Stage 1)**: Light sleep, where heart rate and rhythm may begin to slow.
- **2 (Stage 2)**: Deeper sleep with further slowing of the heart rate.
- **3 (Stage 3)**: Deep sleep, characterized by more stable heart rhythms.
- **4 (Stage 4)**: Another deep sleep phase with stable heart rhythms.
- **R (REM)**: Rapid Eye Movement sleep, often associated with irregular heart rhythms due to dreaming and physiological activity.

These features are crucial for analyzing heart conditions and understanding the relationship between ECG patterns and events/stages in sleep, providing insights into cardiovascular health and sleep disorders.

C. Pre-processing

Butterworth bandpass filtering was applied with a cutoff frequency range of 0.5 to 50 Hz to effectively filter out unwanted noise from the ECG signal. The ECG signals were segmented into overlapping sliding windows, defined by a window size of 30 samples and a 50% overlap.

The variables 'mean', 'variance', 'skewness', 'kurtosis', 'shannon_entropy', 'zero_crossings', 'peak_freq', 'harmonic_ratio', 'heart_rate', initially in object format, were successfully converted to numeric types. The Variables event and sleep_stage columns are encoded into numerical values using the LabelEncoder. Numerical features were scaled using Standard Scaling before fitting the model.

D. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed on the complete dataset prior to splitting it into training and testing subsets.

1) Univariate Analysis: The distributions as shown in the figure 1 shown in these histograms reveal distinct patterns across different features: the mean and variance show sharp, concentrated peaks indicating little variation in these metrics, while skewness exhibits a multimodal distribution suggesting multiple distinct patterns in the data's asymmetry. The kurtosis distribution is heavily right-skewed with multiple peaks at lower values, Shannon entropy shows a roughly normal distribution with some secondary peaks, and zero_crossings has a sharp primary peak with a long right tail. The peak frequency distribution shows multiple distinct peaks suggesting common frequency bands in the data, while the harmonic ratio is extremely concentrated near zero. Most notably, the heart rate distribution appears approximately normal with a slight right skew, centered around 1250-1500, suggesting a relatively consistent range of heart rates in the dataset with some higher outliers.

The distribution of breathing events during sleep as shown in the figure 3 , with a highly imbalanced distribution: 17,400 instances of normal breathing ("no apnea"), only 15 cases of hypopneas (partial breathing interruptions), and 3,115 cases of apneas (complete breathing interruptions). This significant class imbalance, with no apnea being approximately 5.6 times more frequent than apneas and hypopneas combined, suggests that while sleep apnea events are present in the dataset, they represent a minority of the overall sleep breathing patterns.

The distribution of sleep stages as shown in the figure 2 reveals that subjects spent most time either awake (8,127 instances) or in Stage 2 sleep (6,804 instances), which is typical of normal sleep architecture. There's a decreasing prevalence of lighter to deeper sleep stages, with Stage 1 showing 3,462 instances, followed by REM sleep (1,430 instances), Stage 3 (503 instances), and Stage 4 (204 instances) being the least common. This pattern generally aligns with typical sleep architecture, though the high count of awake periods might suggest some sleep disruption.

2) Multivariate Analysis: This pairplot visualization as shown in the figure 4 reveals several interesting patterns and relationships between numerical features in the dataset. The diagonal shows the distribution of individual variables through histograms, while the scatter plots show pairwise relationships between different variables. There appear to be some clear correlations between certain variables as indicated by linear or curved patterns in several scatter plots, while other variable

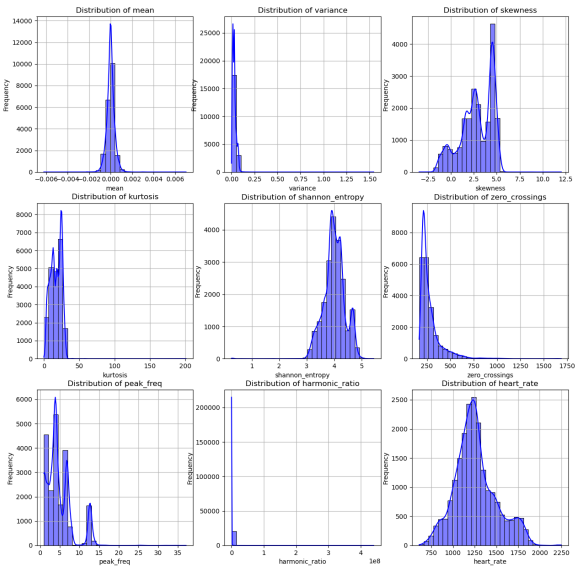

Fig. 1. Distribution of Numerical Variables

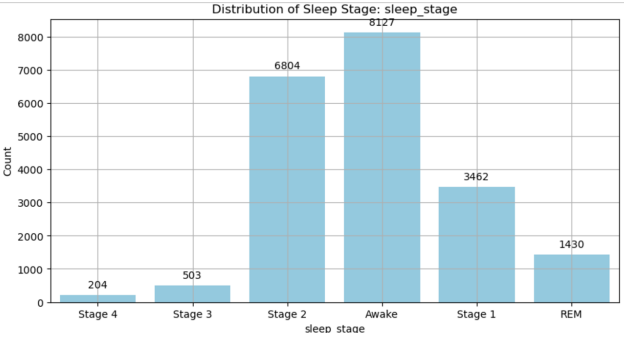

Fig. 2. Distribution of Sleep Stages

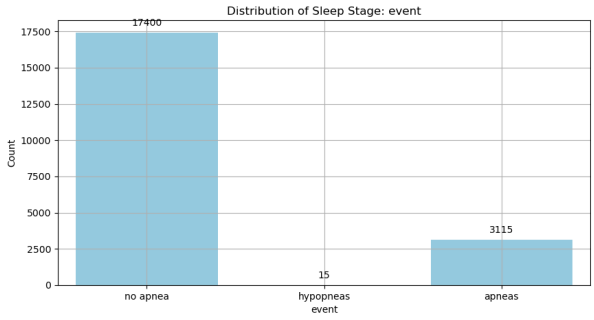

Fig. 3. Distribution of Events

pairs show more dispersed relationships. Some distributions appear to be right-skewed (having a longer tail on the right side), while others show more normal or multimodal distributions. The density of points in various regions of the scatter plots suggests potential clusters or groupings in the data, which could be valuable for understanding underlying patterns or segments in whatever phenomenon this data represents. The plot suggests a rich dataset with complex interactions between variables that would benefit from further statistical analysis to uncover deeper insights.

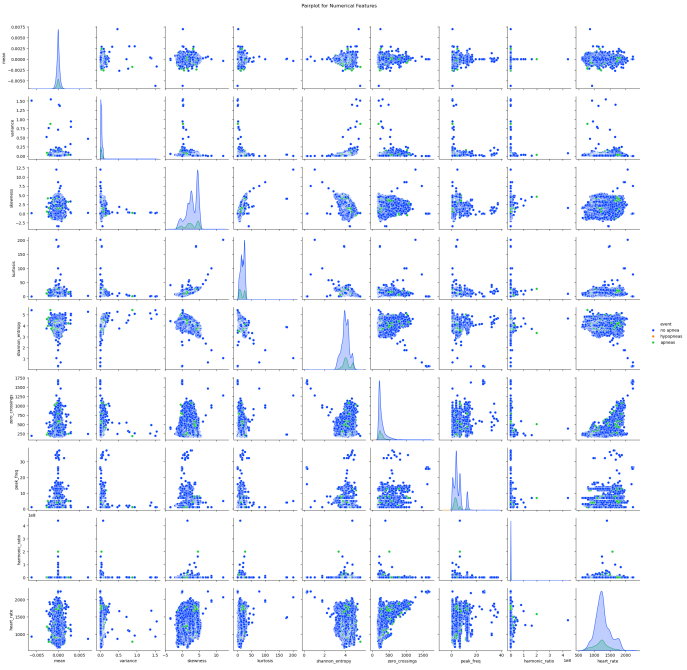

Fig. 4. Multivariate Analysis

E. Correlation Analysis

This correlation matrix as shown in the figure 5 reveals several notable relationships among the features analyzed: there's a strong positive correlation (0.89) between skewness and kurtosis, suggesting similar distributional characteristics, while zero_crossings and heart_rate show a moderate positive correlation (0.64). Shannon entropy exhibits moderate negative correlations with both skewness (-0.40) and kurtosis (-0.51), indicating that as the data becomes more skewed or peaked, its entropy tends to decrease. Most other feature pairs show weak or negligible correlations (values close to 0), with mean and variance showing particularly low correlations with other features, suggesting they capture independent aspects of the data. The presence of both positive and negative correlations, ranging from strong to weak, indicates a complex interplay between these audio-related features, with some measures being closely related while others remain relatively independent.

III. Hidden Markov Model Implementation

A. HMM Model Architecture

The HMM architecture is designed to detect sleep apnea events through the analysis of ECG signals and related data by modeling the sequence of hidden states associated with sleep stages. The model consists of 6 hidden states, each

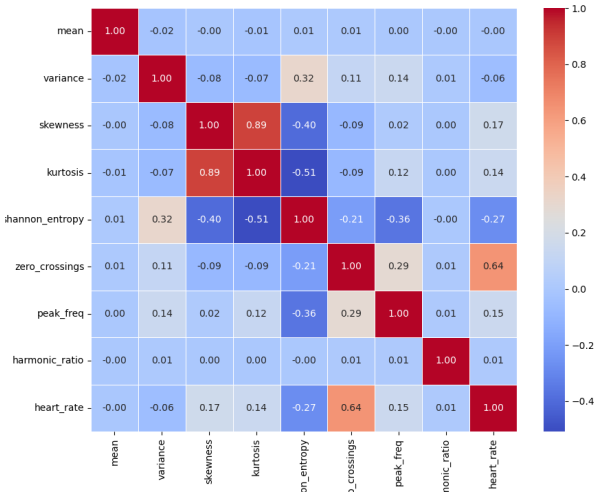

Fig. 5. Correlation Heatmap

representing a different sleep stage, and is built using a Gaussian HMM where each state is modeled as a Gaussian distribution. The architecture assumes ”tied covariance” across the states, which allows the model to treat the covariance of each state as the same, making it suitable for data where different states share similar patterns. The model is trained for up to 2000 iterations with a convergence tolerance of 1e-3, ensuring that it has enough time to optimize the parameters. K-means clustering is used to initialize the model parameters, providing a robust starting point for the optimization process. Additionally, a StandardScaler is applied to normalize the input features, ensuring that each feature has zero mean and unit variance, which enhances model performance and stability. The model also employs LabelEncoders to convert categorical labels for sleep events and stages into numerical values, which are necessary for the HMM to process the data. Finally, the mapping of the hidden states to specific sleep events and stages, enabled the interpretation of the model's predictions in terms of real-world sleep data. This architecture allows the HMM to capture the transitions between sleep stages and apnea events, making it an effective tool for sleep apnea detection.

B. Learning Process

The learning process begins by preparing the dataset for training the Hidden Markov Model (HMM) to detect sleep apnea events. Initially, relevant features, such as mean, variance, skewness, and kurtosis, are extracted from the dataset. The categorical columns, specifically the event and sleep stage data, are converted into numerical values to make them compatible with the model. To improve the model's ability to detect patterns, the feature matrix is normalized, ensuring that all features have a mean of 0 and a standard deviation of 1.

Next, the dataset is split into two parts: one for training and the other for testing, typically using an 80-20% distribution. To address class imbalances, especially in the event labels, synthetic samples are generated for the underrepresented classes. This step ensures the model does not become biased toward the more frequent classes in the dataset.

With the data prepared and balanced, the Hidden Markov Model is trained on the data, learning the underlying patterns and transitions between hidden states, which represent different sleep stages and apnea events. During this training process, the model is able to assign states to the data based on patterns in the observed features. Once the model is trained, a mapping between the hidden states and the corresponding events is created by analyzing how often each event appears within each state.

Finally, the model's performance is assessed based on how well it fits the training data. The model's training score indicates the accuracy of the fit, and the trained model is ready to be tested on the separate test set to evaluate its generalization to unseen data. This entire process enables the HMM to effectively detect sleep apnea events by learning and interpreting the underlying patterns in the data.

C. Evaluation Process

The evaluation process for the Hidden Markov Model (HMM) used in respiration event classification begins with the model making predictions for the test data. These predictions correspond to the hidden states of the model, which are then mapped to specific respiration events. The mapping is based on the most probable event for each predicted state, derived from the model's learned parameters.

Next, a comprehensive performance assessment is conducted using various metrics, including precision, recall, F1-score, and support for each event class. These metrics provide detailed insights into how well the model is identifying each event, and the class labels are displayed according to the encoding used during the training process. Special handling is included for cases where division by zero might occur, such as when a class has no predicted instances, ensuring that such cases are reported as 0 rather than causing errors.

The accuracy of the model is also computed, indicating the overall percentage of correct event classifications. This metric offers a simple yet effective way to measure the general performance of the model.

To further understand the classification performance, a confusion matrix is generated, which compares the true event labels to the predicted event labels. This matrix is visualized as a heatmap, providing a clear view of the correct and incorrect classifications made by the model.

Finally, the predicted states and corresponding events are returned, allowing for further analysis or evaluation of the model's performance. This process helps identify areas where the model may need improvement and guides the next steps for model refinement.

D. Decoding Process

The decoding process in the context of evaluating sleep stage predictions using a Hidden Markov Model (HMM)

involves several steps that together provide a comprehensive view of the model's performance.

The process starts with the application of the Viterbi algorithm, which is used to decode the most likely sequence of hidden states (in this case, the sleep stages) for a given set of input features (the test data). The Viterbi algorithm finds the optimal sequence of states by maximizing the likelihood of the observed data, based on the model's parameters, such as transition and emission probabilities. This results in the sequence of predicted sleep stages, which is considered the best estimate of the sleep stage progression over time.

Following the prediction of sleep stages, the model's performance is evaluated using a variety of metrics. The classification report provides key performance metrics such as precision, recall, F1-score, and support for each sleep stage class, offering a detailed breakdown of how well the model performed for each stage. These metrics help assess the model's ability to correctly classify each sleep stage while accounting for potential misclassifications.

The accuracy of the model is computed by comparing the predicted sleep stages with the true stages in the test set, providing an overall measure of the model's performance. Additionally, a confusion matrix is generated, which visualizes the true versus predicted sleep stages. This matrix highlights where the model made correct predictions and where it misclassified sleep stages, further aiding in performance analysis.

Another important aspect of the decoding process is the examination of the transition matrix, which captures the probabilities of transitioning from one sleep stage to another. This matrix provides insight into how the model understands the temporal dependencies between consecutive sleep stages, which is a key feature of sleep patterns. By visualizing the transition matrix, it is possible to interpret how likely the model is to move between different stages, such as from light sleep to deep sleep or from REM to wakefulness.

Finally, the sequence of predicted sleep stages is returned, providing the decoded sleep stages for the test data. This output can be further analyzed for model validation or used in downstream tasks. Through this decoding and evaluation process, the model's ability to accurately predict sleep stages can be assessed, and areas for improvement can be identified.

IV. Result

The Hidden Markov Model (HMM) trained for respiration event classification and sleep stage decoding demonstrates mixed performance.

A. Respiration Events

The model achieved a classification accuracy of 51.92% as shown in the figure 6, performing slightly better than random guessing. While it predicts the "No Apnea" class well, with 86% precision, its recall for this class is lower at 52%, indicating missed instances. For "Apneas," the model achieved a recall of 50% but a precision of only 22%, suggesting frequent false positives. The "Hypopneas" class had extremely poor performance, with just 1% precision despite 83% recall, signifying significant misclassification. The weighted average metrics (precision: 76%, recall: 52%, F1-score: 60%) reveal that the model is biased toward the dominant "No Apnea" class, while minority classes require improved feature extraction and class balancing. Confusion matrix analysis as shown in the figure 7 highlights overlapping features between "Apneas" and "No Apnea" and severe challenges in identifying "Hypopneas."

Respiration Event Classification Report:

	precision	recall	f1-score	support
apneas	0.22	0.50	0.31	649
hypopneas	0.01	0.83	0.02	6
no apnea	0.86	0.52	0.65	3451
accuracy			0.52	4106
macro avg	0.37	0.62	0.33	4106
weighted avg	0.76	0.52	0.60	4106

Event Classification Accuracy: 51.92%

Fig. 6. Classification Report of Respiration Events

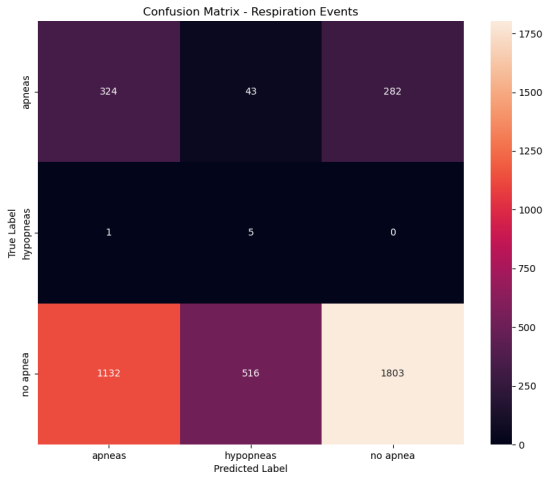

Fig. 7. Confusion Matrix of Respiration Events

B. Sleep Stages

Sleep stage decoding accuracy was 11.40% as shown in the figure 8, reflecting considerable difficulty in predicting stages accurately. The model performed best for "Awake" (36% precision, 22% recall) and "Stage 2" (better classification), while other stages like REM, Stage 1, Stage 3, and Stage 4 showed very low precision and recall, indicating misclassification and poor generalization. The confusion matrix as shown in the figure 9 highlights that the model correctly classified 367 instances for the "Awake" stage and 420 for "Stage 2," demonstrating relatively better performance for these stages. However, REM, Stage 1, Stage 3, and Stage 4

exhibit poor classification accuracy with significant overlap between stages, indicating high misclassification rates. State transition analysis as shown in the figure 10 revealed uneven probabilities, with notable transitions from ”Awake → Stage 4” (49%) and ”REM → Stage 3” (61%). Stable states like Stage 3 and Stage 4 exhibited high self-transition probabilities, but overall imbalanced transitions suggest the need to better model time-series dependencies to enhance classification.

Sleep Stage Classification Report:

	precision	recall	f1-score	support
Awake	0.36	0.22	0.28	1658
REM	0.04	0.02	0.03	291
Stage 1	0.11	0.07	0.09	689
Stage 2	0.11	0.01	0.02	1325
Stage 3	0.04	0.18	0.07	99
Stage 4	0.01	0.36	0.02	44
accuracy			0.11	4106
macro avg	0.11	0.14	0.08	4106
weighted avg	0.21	0.11	0.13	4106

Sleep Stage Classification Accuracy: 11.40%

Fig. 8. Classification Report of Sleep Stages

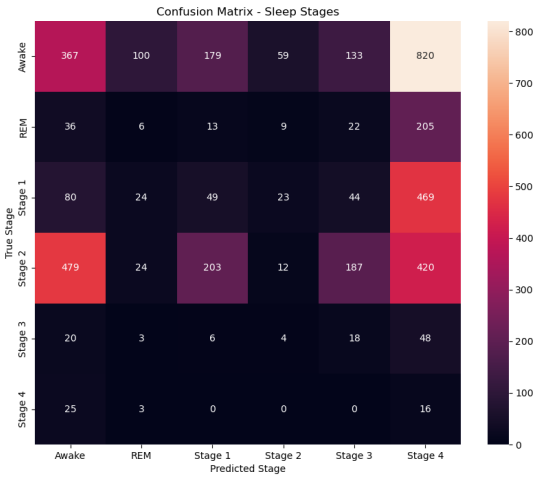

Fig. 9. Confusion Matrix of Sleep Stages

V. Discussion

The study demonstrates the application of Hidden Markov Models (HMM) for classifying respiration events and sleep stages to aid in sleep apnea detection. While the model achieved a modest accuracy of 51.92% for respiration events, its performance highlighted significant challenges, particularly in handling imbalanced datasets. The detection of dominant events like ”no apnea” was relatively accurate, but rare events such as ”hypopneas” were poorly classified, reflecting the limitations of the model in addressing minority classes. Similarly, the sleep stage decoding accuracy of 11.40% revealed

Fig. 10. State Transition Probabilities

difficulties in distinguishing complex physiological patterns, indicating the need for richer feature sets beyond the statistical and frequency-domain measures used.

Dataset preparation also presented challenges, as the windowing approach and manual selection of dominant events might have oversimplified transitions and overlapping signals. Additionally, the Gaussian assumptions of the HMM limited its ability to model nonlinear relationships, as seen in the transition matrix patterns. Despite these challenges, the study highlights the feasibility of using machine learning for automated sleep apnea detection while emphasizing the need for advanced models, multimodal data integration, and improved preprocessing to enhance performance and generalizability.

VI. Conclusion

The model demonstrates basic capability in classifying respiration events with moderate accuracy but struggles with imbalanced classes and complex dynamics, as evident in the low performance for rare events like hypopneas and sleep stages. The significant gap between training success and evaluation results suggests overfitting, while the poor sleep stage classification highlights the need for more robust features and advanced modeling techniques. Improvements in data balancing, feature engineering, and model architecture are essential to enhance its clinical applicability and overall predictive performance.

VII. Limitations

The current model faces several limitations that impact its performance and clinical utility. One key challenge is the imbalanced dataset, where rare events like hypopneas and underrepresented sleep stages such as Stage 3 and REM are poorly predicted. This is compounded by the model’s inability to capture complex physiological relationships, as the Gaussian HMM, though effective for basic temporal patterns,

lacks the sophistication needed for intricate sleep dynamics. The low precision and recall across most sleep stages suggest that the feature set, which includes statistical and frequency-domain measures, may not fully encapsulate the nuanced variations in sleep physiology. Additionally, overfitting during training indicates that the model is not generalizing well to new data, limiting its effectiveness for clinical application.

In preparing the data, limitations also arise. The process of parsing sleep study annotation files to categorize sleep stages and apnea events introduces potential inaccuracies, especially when dealing with missing or unknown stages. While post-processing attempts to fill gaps in stage information, it may still lead to misclassifications or incomplete stage transitions. Furthermore, the approach of dividing data into overlapping windows and selecting the most frequent event and stage within each window can result in oversimplification, as this method may fail to capture subtle shifts in sleep patterns and apnea events. These data preparation challenges further affect the model's ability to reliably classify events and stages.

VIII. Future Scope

Future improvements can address these challenges through multiple avenues. Incorporating advanced feature extraction techniques, such as deep neural network embeddings, can provide richer representations of the physiological data. Exploring hybrid models that combine HMMs with deep learning architectures, such as LSTMs or GRUs, may enhance the model's capacity to capture temporal dependencies and non-linear relationships. Addressing data imbalance through advanced augmentation techniques, like GAN-generated synthetic samples, can improve rare event classification. Furthermore, fine-tuning the HMM's architecture, experimenting with different covariance types, and optimizing hyperparameters can lead to better performance. Incorporating multimodal data, such as EEG and respiratory signals, could yield a more holistic understanding of sleep patterns and improve classification accuracy.

IX. Generalization

For broader applicability, the model requires validation on larger, more diverse datasets spanning various demographics and clinical conditions. Integration of multimodal data, such as EEG, ECG, and respiratory signals, may improve robustness and ensure the model performs reliably in real-world settings. For real-world deployment, the model must demonstrate robust generalization across diverse populations, including different age groups, genders, and clinical conditions. This can be achieved by training and validating the model on larger, more diverse datasets sourced from multiple centers. Regularization techniques, cross-validation, and domain adaptation methods can ensure better performance across unseen data. By integrating multimodal data and addressing dataset heterogeneity, the model could achieve clinical-grade reliability, paving the way for its use in sleep studies, wearable devices, and personalized sleep therapy applications.

References

[1] Y. Ichimaru and G. B. Moody. Development of the polysomnographic database on cd-rom. *Psychiatry and Clinical Neurosciences*, 53:175–177, April 1999.