Estimating and Analyzing Sleep Stages Through Wearable Sleep Tracker

[1]Mehmet Furkan Çalışkan , mfcaliskan03@gmail.com
[1]Ufuk Cefaker , ufukcefaker@gmail.com
[1],
[1,]Tugba Gurgen Erdogan, tugba@cs.hacettepe.edu.tr
[1] Hacettepe University, Computer Engineering Department, Software Engineering Research Group, Ankara, Turkey

Abstract

Context:
Obstructive sleep apnea (OSA) is a prevalent sleep disorder causing apnea and hypopnea episodes, linked to cardiovascular diseases and hypertension. Polysomnography (PSG) is effective but costly and requires specialized equipment. This project uses the Dreamt dataset to analyze physiological data like electrodermal activity and heart rate variability for sleep stage detection via machine learning, while studying the impact of conditions like diabetes and hypertension on sleep apnea severity. The aim is to create accessible, affordable tools for smart devices.

Purpose:
This project seeks to overcome traditional diagnostic limitations for sleep apnea by using machine learning to detect sleep stages and apnea severity through physiological data. It explores links between markers and conditions like diabetes and hypertension to develop cost-effective sleep monitoring tools and better understand sleep apnea factors.

Method:
Machine learning models, including XGBoost, Random Forest, SGDClassifier, and Deep Learning, were applied to the Dreamt dataset. Physiological data such as electrodermal activity and heart rate variability were analyzed. Matplotlib was used for visualizing correlations between sleep apnea and comorbidities like diabetes and hypertension.

Results:
Machine learning models effectively detected sleep stages and highlighted correlations between sleep apnea and comorbidities. Key findings and model performances are presented in the "Experiment Results" section.

Keywords:
OSA, apnea, hypopnea, PSG, EEG, EOG, Sleep stages, Machine Learning, Deep Learning, Sleep disorders, EDA, BVP, HR, BMI, OAHI

1. INTRODUCTION

Obstructive sleep apnea (OSA) syndrome is a condition characterized by the collapse of the upper airway during sleep, leading to frequent and repetitive episodes of apnea (cessation of breathing) and

hypopnea (shallow breathing). These disruptions result in significantly negative effects on health by causing high negative intrathoracic pressure, recurrent hypoxia/hypercapnia [1], and frequent arousals from sleep. OSA triggers heightened sympathetic nerve activity, oxidative stress, systemic inflammation, and vascular endothelial dysfunction. Moreover, it is strongly linked to serious health complications such as hypertension, cardiovascular diseases, and cerebrovascular disorders, which substantially elevate the risk of mortality. [2-5]

Sleep apnea, a prevalent sleep disorder globally and in Turkey, poses significant health challenges. The prevalence of obstructive sleep apnea syndrome (OSAS) is reported to be 3.9% in men and 1.2% in women worldwide, highlighting its widespread impact. [6]

Polysomnography (PSG) is a comprehensive diagnostic tool for assessing sleep disorders by monitoring various physiological parameters during sleep. The cost and duration of PSG vary by region and healthcare system. In the United States, an in-laboratory PSG typically costs $1,000 to $5,000, while home-based studies are more affordable, ranging from $300 to $600, depending on insurance and provider. The study is conducted overnight, lasting 6 to 8 hours to capture a full sleep cycle, with about an hour of preparation. Results analysis and interpretation may take several days, influenced by the facility's workload and resources. [7]

The advent of wearable technology and datasets like Dreamt has revolutionized sleep analysis by providing physiological data—such as electrodermal activity, accelerometry, and heart rate variability—instead of relying solely on traditional EEG and EOG measures. Despite this progress, accessible and effective sleep stage detection remains elusive, particularly for individuals with sleep disorders, due to limitations in current methods and datasets. Our project seeks to address this gap by leveraging the Dreamt dataset to enhance sleep stage detection using physiological signals, aiming to eliminate the dependency on EEG and EOG. Through advanced machine learning models and feature engineering, we aim to explore correlations between physiological markers and sleep disorders, enabling affordable, user-friendly sleep tracking solutions for everyday devices like smartwatches. This paper introduces our pipeline for data preprocessing, feature extraction, and classification, alongside an evaluation framework, setting the stage for practical tools in both medical and consumer health applications.

The remainder of this article is organized as follows. Section 2 provides an overview of related studies and highlights the need for this research with respect to related studies. Section 3 explains the methodology employed while carrying out this study... Section 4 provides the results in correspondence with the research questions. Section 5, we provide overall conclusions and plans for future work.

2. BACKGROUND AND RELATED WORK

Table 1: Summary of the related work

Year [ref], Venue	Title	Objective	Datasets	Findings w.r.t. ....
Wang et al. [9] , 2024, (Conference)	**Addressing wearable sleep tracking inequity: a new dataset and novel methods for a population with sleep disorders**	**Using the dataset from this paper to train model and finding correlation between past diseases and OSA(Obstructive Sleep Apnea).**	**https://physionet.org/content/dreamt/1.0.1/data/#files-panel**	**Without using traditional EEG and EOG measures, predicting OSA with respect to electrodermal activity, accelerometry, and heart rate variability**

Nochino et al. [11], 2019, Biomedical Engineering Letters (Journal)	**Sleep stage estimation method using a camera for home use**	**We are using this paper, to understand if the body acceleration and sleep stage is correlated.**	**N/A**	**Body movement correlates with sleep stages, and accelerometry provides a non-invasive, low-cost method for their estimation.**
Giles et al. [1], 2006, Cochrane Database of Systematic Reviews (Journal)	**Continuous positive airways pressure for obstructive sleep apnoea in adults**	**This paper evaluates the effectiveness of CPAP therapy for treating OSA in adults, focusing on sleep quality and health outcomes.**	**N/A**	**CPAP therapy effectively improves sleep quality, reduces apneic episodes, and lowers cardiovascular risks in managing OSA.**
Peppard et al. [2], 2000, New England Journal of Medicine (Journal)	**Prospective Study of the Association between Sleep-Disordered Breathing and Hypertension**	**It explores how sleep-disordered breathing contributes to the development of hypertension over time in adults.**	**N/A**	**Sleep-disordered breathing raises hypertension risk, highlighting the importance of early diagnosis and treatment.**
Penzel et al. [8], 2003, Neuropsychopharmacology (Journal)	**Dynamics of heart rate and sleep stages in normals and patients with sleep apnea.**	**To understand the relation between heart rate variability between stages, and the relation of sleep stages and sleep disorders.**	**N/A**	**HRV decreases with deeper sleep but rises in REM, with healthy individuals spending most time in light sleep, while these patterns vary in sleep disorders.**
Gaiduk et al. [10], 2022, IEEE Journal of Biomedical and Health Informatics ()Journal	**Estimation of sleep stages analyzing respiratory and movement signals**	**Deriving useful features and estimating Wakefulness, REM and Non-REM stages from respiratory and movement signals.**	**N/A**	**Movements during sleep is correlated and can be used to estimate sleep stages and wakefulness.**

Summary of Related Studies

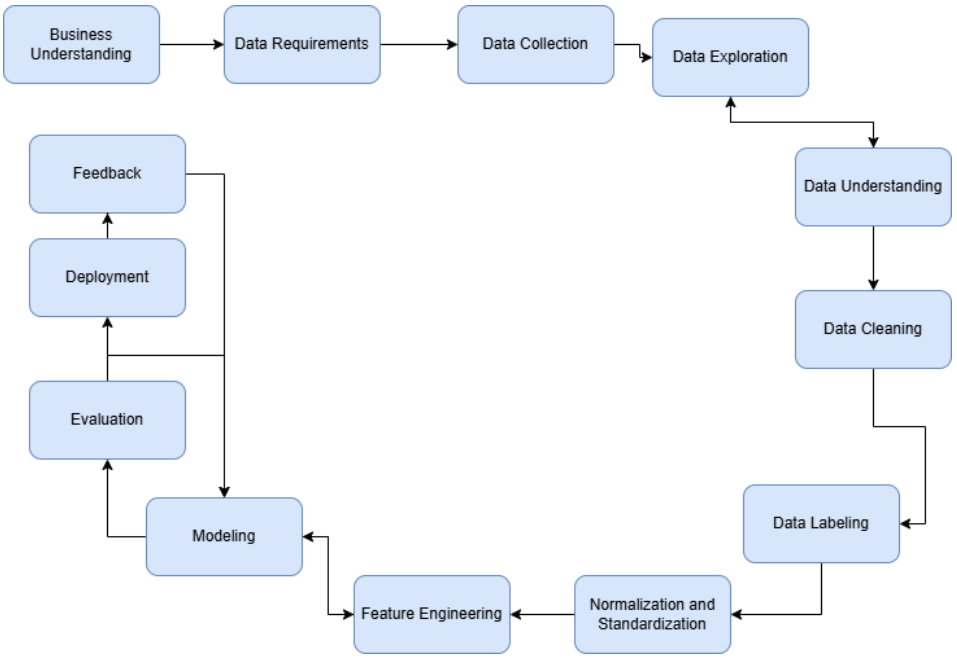
Wang et al. (2024) introduced a dataset enabling OSA detection using non-invasive signals like electrodermal activity, accelerometry, and heart rate variability, forming the foundation for our work. Other studies, such as Nochino et al. (2019), highlighted accelerometry's role in sleep stage estimation, while Giles et al. (2006) and Peppard et al. (2000) emphasized the clinical importance of OSA

detection and its link to health outcomes. These studies guide our approach in developing non-invasive, comprehensive sleep monitoring solutions.

Our Contributions

We extend these works by using the DreamT dataset to train a machine learning model capable of predicting both OSA and all sleep stages, addressing accessibility challenges and advancing non-invasive sleep analysis.

3. METHODOLOGY



3.1 Business requirements:

Goal:
The overarching goal of our project is to improve the understanding and detection of sleep-related disorders and sleep stages using physiological data. Specifically:

Correlation Analysis: Investigate relationships between patient medical history, demographic data, and sleep disorders with specific apnea events (Obstructive Apnea, Central Apnea, Hypopnea).

Sleep Stage Detection: Develop a machine learning model capable of detecting sleep stages using physiological data, reducing reliance on traditional EEG/EOG methods.

Scope:

1. **Data Exploration and Correlation Analysis**:

   Utilize the ***participant_info*** dataset to analyze relationships between medical and demographic factors (e.g., AGE, GENDER, BMI, MEDICAL_HISTORY, Sleep_Disorders) and sleep disorder events such as Obstructive Apnea, Central Apnea, and Hypopnea, identifying patterns or significant predictors for these events.

2. **Model Development for Sleep Stage Detection**:

   For sleep stage classification, preprocess data by handling missing values and normalizing metrics, extract features like derivatives and statistical measures, and use dimensionality reduction (e.g., PCA). Train and evaluate models (e.g., Random Forest, SVM) using cross-validation and metrics like accuracy and F1-score.

3.2 Research Questions

1. **RQ1. Relationship Analysis**:

   RQ1.1. What is the relationship between ages and the time spent in sleep stages? (3.5.7.)

   RQ1.2. How do patterns in measurements relate to sleep stages, and what insights can they provide about patients during sleep? (3.5.7)

   RQ1.3 How does a patient's medical history influence the occurrence of the most prevalent apnea events? (4.2.3)

2. **RQ2. Sleep Stage Detection**:

   RQ2.1 Can physiological signals like EDA, BVP, and HR effectively distinguish between sleep and wake states without relying on traditional EEG/EOG measures? (4.2.3)

   RQ2.2 Which physiological signals and machine learning models are most effective in accurately estimating sleep stages, and how do they compare in performance? (4.2.2)

   RQ2.3 What is the relationship between sleep apnea and associated conditions like excessive daytime sleepiness (EDS), depression, bruxism, and snoring, compared to common comorbidities like hypertension? (4.2.3)

   RQ2.4 What is the relationship between patients' transitions through sleep stages and their likelihood of entering the W state? (4.2.4)

3.3 Data requirements:

Participant Information Dataset:

This dataset provides demographic, clinical, and sleep-related attributes that are essential for the initial correlation analysis. The key variables required include:

Column Name	Description	Data Type
SID	A unique identifier for each participant.	Categorical
AGE	Numerical data representing the participant's age.	Numerical
GENDER	Categorical data denoting gender.	Categorical
BMI	Numerical data reflecting the participant's Body Mass Index.	Numerical
OAHI	Obstructive Apnea-Hypopnea Index, a numerical metric quantifying apnea events.	Numerical
AHI	Apnea-Hypopnea Index, a summary measure of both apnea and hypopnea events.	Numerical
Mean_SaO2	Numerical data indicating the mean oxygen saturation level during sleep.	Numerical
Arousal Index	A numerical value representing the frequency of sleep disturbances.	Numerical
MEDICAL_HISTORY	Categorical or multi-label data capturing patient health conditions.	Categorical/Multi-Label
Sleep_Disorders	Categorical data indicating the presence or type of sleep disorders.	Categorical

To ensure data quality, the dataset must be free of duplicate entries, contain consistent units (e.g., BMI in kg/m², SaO2 as a percentage), and have uniform categorical labels.

Patient Data Dataset:

This dataset comprises physiological measurements and annotations that form the basis for machine learning-based sleep stage detection. The essential variables include:

Column Name	Description	Data Type
TIMESTAMP	A consistent and well-formatted record of data collection times (e.g., ISO 8601).	Numerical
BVP (Blood Volume Pulse)	Numerical data representing cardiovascular activity.	Numerical
ACC_X, ACC_Y, ACC_Z	Numerical data from accelerometer readings that capture body movements.	Numerical
TEMP	Numerical data for skin temperature.	Numerical
EDA (Electrodermal Activity)	Numerical data reflecting sympathetic nervous system activity.	Numerical
HR (Heart Rate)	Numerical data derived from the BVP signal.	Numerical
IBI (Interbeat Interval)	Numerical data representing the time intervals between consecutive heartbeats.	Numerical
Sleep_Stage	Categorical data denoting sleep stages (e.g., Wake, N1, N2, N3, REM).	Categorical
Obstructive_Apnea	Binary annotation for obstructive apnea events.	Binary
Central_Apnea	Binary annotation for central apnea events.	Binary
Hypopnea	Binary annotation for hypopnea events.	Binary
Multiple_Events	Binary annotation for multiple respiratory events.	Binary

3.4 Model requirements:

Feature Selection:
The features in this study are divided into two groups: physiological features and demographic/clinical features. Physiological features include Blood Volume Pulse (BVP), Heart Rate (HR), Electrodermal Activity (EDA), Skin Temperature (TEMP), accelerometer readings (ACC_X, ACC_Y, ACC_Z), and Interbeat Interval (IBI). Demographic and clinical features encompass age, gender, Body Mass Index (BMI), medical history (e.g., pre-existing conditions, comorbidities), and sleep disorders represented as binary or categorical labels. Additional clinical metrics include Oxygen Saturation (Mean_SaO2), Arousal Index, Apnea-Hypopnea Index (AHI), and Obstructive Apnea-Hypopnea Index (OAHI). The target variables are Obstructive Apnea, Central Apnea, and Hypopnea for correlation analysis, and Sleep Stages (Wake, N1, N2, N3, REM) for sleep stage detection. Engineered features include like Heart Rate Scaled By Mean, Heart Rate Scaled By Median, and Abnormal_IBI_Moving_Average_640.

Model Types:

Model 1. Classical Machine Learning Models:

> **Model 1.1. Random Forest:** Robust for tabular data and useful for analyzing feature importance.
>
> **Model 1.2. Stochastic Gradient Descent (SGD):** A scalable and efficient algorithm for large datasets, often used for linear classification.
>
> **Model 1.3. Gradient Boosting Methods (e.g., XGBoost):** High-performing models that excel in classification tasks with structured data.
>
> **Model 1.4. Logistic Regression:** A simple yet effective model, offering a balance of performance and interpretability.

Model 2. Deep Learning Neural Network: Used to capture complex patterns in data, though they may struggle with imbalanced classes without appropriate tuning.

Evaluation Criteria:

Performance will be evaluated using key metrics: accuracy for overall correctness, F1-score to balance precision and recall on imbalanced datasets, and precision/recall to assess minority class performance. A confusion matrix will analyze predictions across sleep stages, while logarithmic loss will measure the confidence of probabilistic predictions.

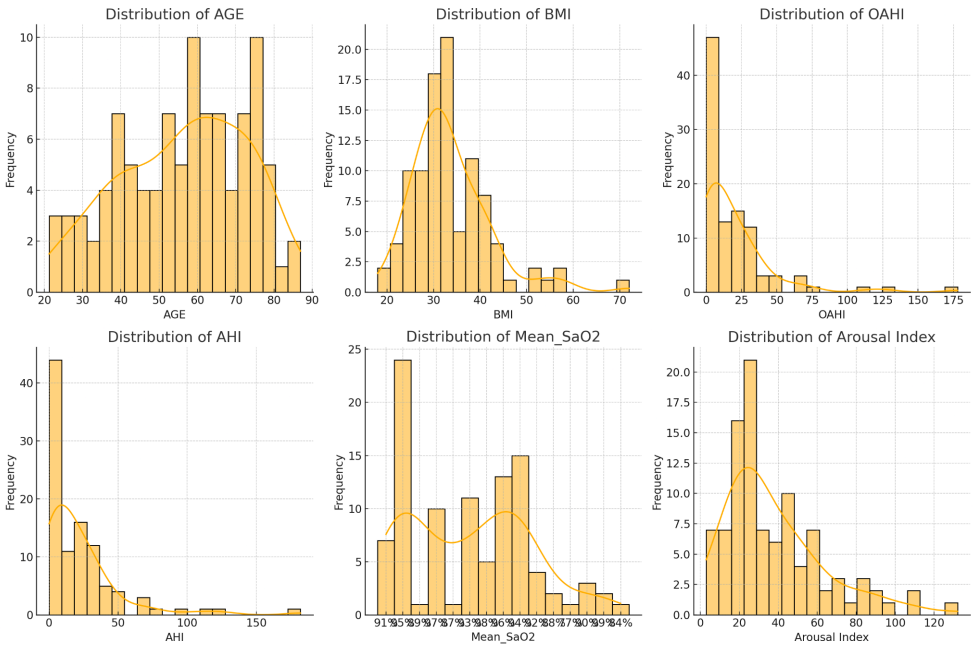
3.5 Data Exploration and Understanding:

3.5.1 Dataset Overview:

The data for this study was sourced from the **DREAMT: Dataset for Real-time Sleep Stage Estimation using Multisensor Wearable Technology**. This dataset provides physiological, demographic, and clinical data collected from 100 patients. The primary aim of the dataset is to enable sleep stage classification and sleep disorder analysis using non-invasive wearable technology.
The dataset includes rich physiological signals such as blood volume pulse, accelerometer readings, and electrodermal activity, in addition to patient demographic and clinical information. These data points are critical for exploring relationships between medical history, sleep disorders, and apnea events, as well as for developing machine learning models for sleep stage detection.

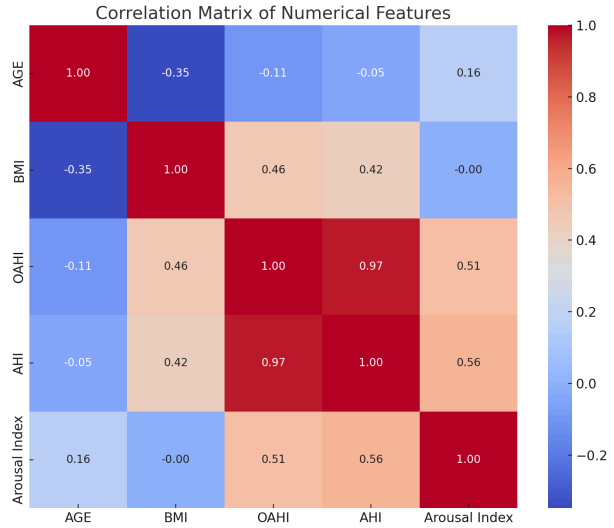
3.5.2 Visualizations

1. **Distribution of Key Demographic and Clinical Features**



The distributions show that patient ages peak around middle-aged and older adults, BMI centers between 25–35, and OAHI and AHI are skewed towards lower values, indicating mild apnea-hypopnea in most cases. Mean SaO2 clusters around 90–95%, while the Arousal Index shows moderate values with fewer high occurrences, reflecting variability in clinical severity.

2. **Correlation Heatmap of Numerical Features**



A heatmap was generated to explore relationships between numerical variables such as AGE, BMI, OAHI, AHI, Mean_SaO2, and Arousal Index. Strong correlations, such as those between AHI and OAHI, highlight related factors.

3. **Time-Series Example of Physiological Data**

Physiological signals such as Heart Rate (HR) and Electrodermal Activity (EDA) show variations across the night, aligning with sleep stages and apnea events.

3.5.3 Data Cleaning:

The DREAMT dataset presented minimal missing values in the recorded data, which greatly simplified the preprocessing phase. However, certain nuances required careful attention, particularly regarding the **Inter-Beat Interval (IBI)**, which was recorded with a slight delay after the onset of the Polysomnography (PSG) process. This initial delay created gaps that posed challenges during feature engineering. While no missing values were present after the PSG officially began, the gaps during the preparatory phase were relevant for establishing a baseline for wakefulness. To address this, we applied a **back-filling technique** to fill these initial missing seconds, ensuring numerical stability during subsequent feature calculations.

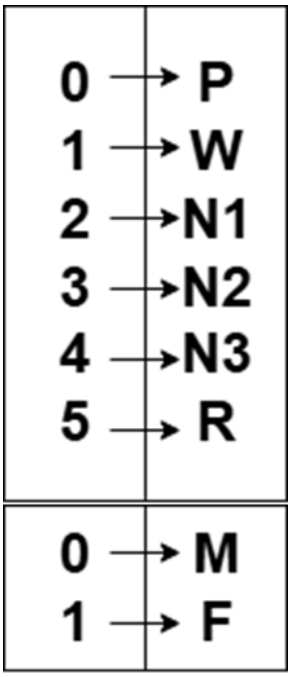
For all other recorded columns, no missing values were detected, which streamlined this aspect of the preprocessing.

A more challenging issue arose with **sleep stage labels**, which were occasionally missing for some patients. These gaps typically occurred when the PSG setup produced unreliable or incomplete data. According to Wang et al. (2024), two patients had missing labels for approximately 15 minutes each, while the remaining four patients experienced brief gaps of 30 seconds or less during the initial setup phase. Given the significant proportion of missing data, the two patients with the longest label gaps were excluded from further analysis. For the remaining patients, missing labels were filled using the subsequent sleep stage label to maintain the temporal consistency of the data.

3.5.4 Data Labeling:

Another critical preprocessing step involved handling missing values in columns used for **apnea labeling**. These columns recorded values as either 1 (indicating apnea events) or NaN. We replaced the NaN entries with 0 to indicate the absence of apnea events, thereby maintaining consistency in the dataset for classification purposes.

We also encoded the sleep stages and genders with numerical labels as shown in the figure.



3.5.6 Data Normalization and Standardization:

To enhance the model's understanding, we incorporated additional patient information, such as their **medical history**. Disorders and conditions were standardized and transformed using **one-hot encoding** to ensure categorical data could be utilized effectively by the model. Moreover, metrics that were not derived from the sleep study, such as **BMI, age, and gender**, were **normalized** to ensure a balanced training process and to prevent these values from dominating other features in the dataset.

We applied normalization to the sleep recording data across all patients using **min-max normalization**. This method was chosen because it accounts for abnormal data points, which can be particularly explanatory in this dataset. For instance, spikes in heart rate during nightmares might provide valuable insights. IBI data is excluded from the normalization process since the data is already mostly between 0 and 1.

Instead of normalizing each patient's data individually, we calculated the normalization parameters (minimum and maximum values) across the entire patient population. This

approach maintained the relationship between each individual's data changes while preserving the dataset's overall integrity.

This preprocessing approach ensured that the dataset was robust, consistent, and well-prepared for the feature engineering and analysis stages, while addressing the challenges posed by missing values thoughtfully and systematically.

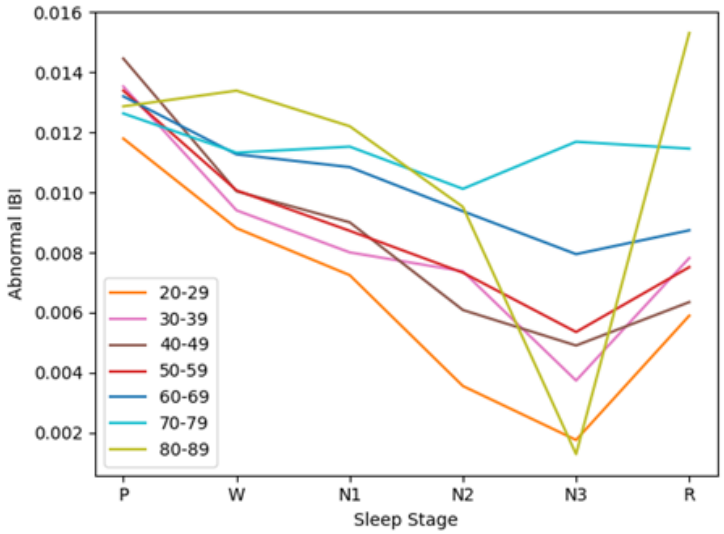
3.5.7 Feature Engineering:

ACC:

To better understand bodily movements during sleep, we combined acceleration data from all three dimensions into a single 3-dimensional acceleration metric. By summing the accelerations across each axis, this feature gives us a clearer picture of overall movement patterns, providing valuable insights into how the body shifts during different sleep stages.

Abnormal_IBI_Moving_Average_640

This feature captures how much the time between consecutive heartbeats (Inter-Beat Interval, or IBI)



varies across sleep stages. To calculate it, we analyzed each patient's data in 10-second windows, computing both the moving mean and moving standard deviation within each window. Every 10 seconds adds 640 rows to the dataset. When an IBI deviates by more than four moving standard deviations from the moving mean, it's flagged as an abnormal IBI (labeled as "1"). We then examined how frequently these abnormalities occurred across different sleep stages.

In the accompanying figure, you can see how this metric relates to sleep stages across age groups. It shows a distinct pattern, decreasing until the N3 stage and rising again during REM sleep. However, because the scale of this metric is very small, we normalized it by dividing each value by its respective mean. This scaling ensures consistency and improves model performance by bringing the metric into

a comparable range. HRV is the measure we have t
correl
deepn
REM s

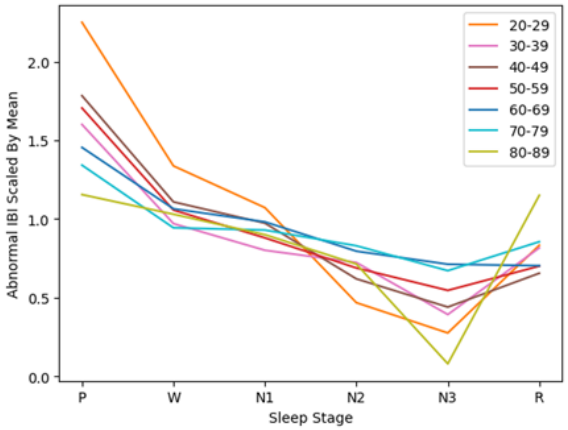


Figure. Abnormal IBI counts scaled by mean

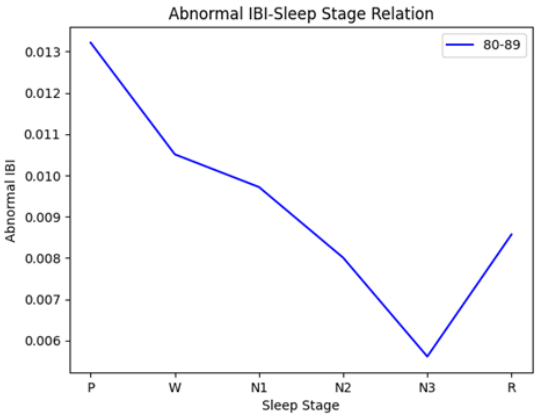
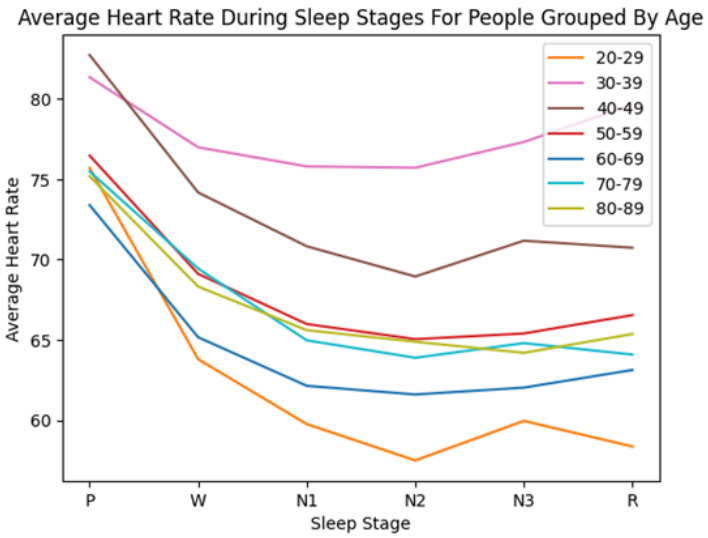
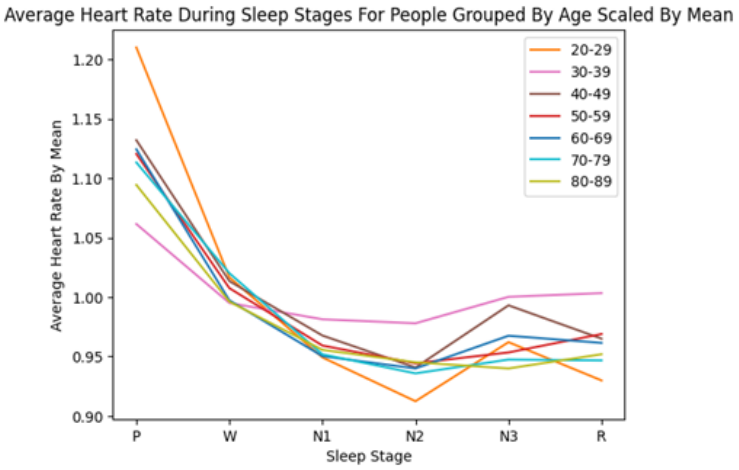


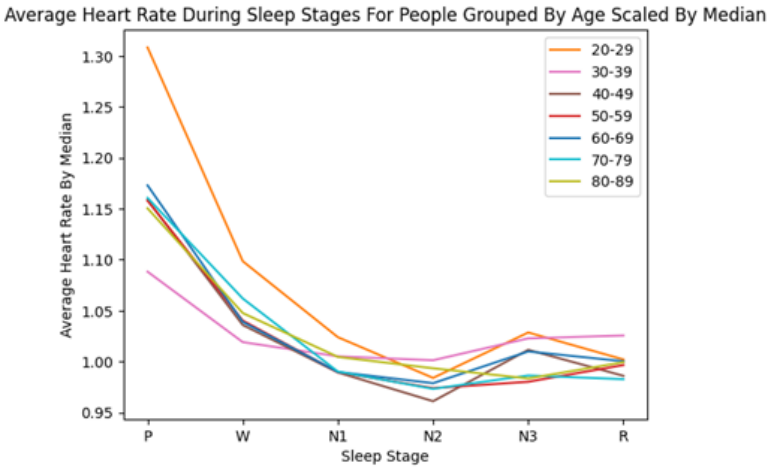
Figure. Abnormal IBI counts for every 10 seconds

Heart Rate Scaled By Mean And Median

We noticed an intriguing connection between heart rate and sleep stages, but this relationship seems to shift with age. During the deeper stages of sleep, heart rates generally slow down. However, in middle-aged individuals, this slowdown isn't as pronounced as it is in older or younger people. While this variation could stem from personal differences, we aimed for a more consistent metric that the model could learn from more easily. To achieve this, we normalized heart rates by dividing each person's heart rate by their own mean and median values. This approach helps standardize the data and makes the model's learning process smoother.







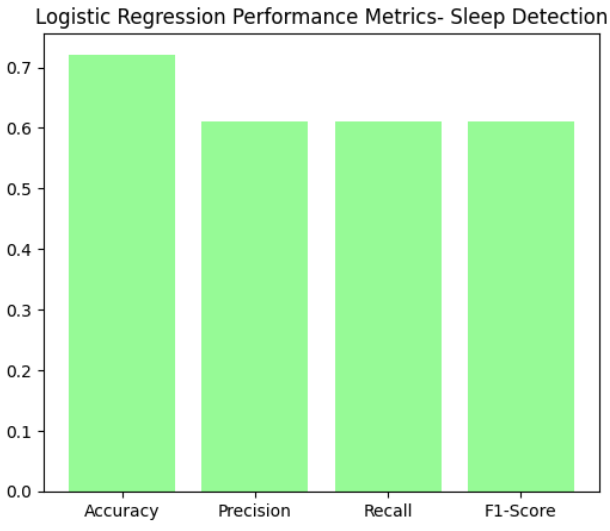
3.5.8 Outcome: Development of a Data Pipeline

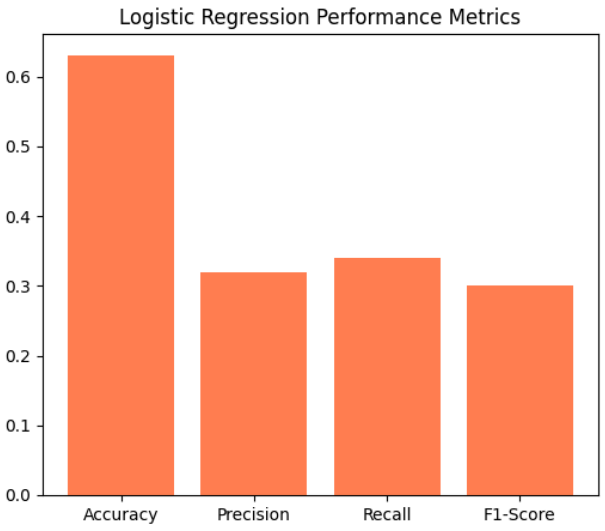
As a result of our efforts, we successfully developed a robust data pipeline. This pipeline streamlines the process of collecting, processing, and analyzing physiological data, ensuring seamless integration and scalability. It enables efficient feature extraction, normalization, and analysis, making it easier to uncover meaningful patterns and relationships within the data. This foundational infrastructure is critical for advancing our research and building more accurate and reliable models for sleep stage detection.

Based on our evaluation, we chose Logistic Regression as the primary model due to its consistent and balanced performance across all metrics, including precision, recall, and F1-score. This selection aligns with the model's ability to handle class imbalance effectively and provide interpretable results.

Outcome: The classification results demonstrate that Logistic Regression outperforms other models, with clear and meaningful visualizations that highlight its ability to predict multiple classes effectively.

Model Evaluation: Through detailed analysis of accuracy, precision, recall, and F1-score, Logistic Regression emerged as the most reliable and well-suited algorithm for this task.





4. EXPERIMENTS

4.1 Experimental Setup

We specifically chose to work with a single dataset to ensure consistency and depth in our analysis. This dataset is chosen for its accessibility and cost-effectiveness, making it an ideal choice for practical applications. Its features are designed to be easily and inexpensively implemented, enabling broad usability without the need for specialized equipment or complex setups. This accessibility aims that advanced sleep analysis methods can be democratized, reaching a wider audience. By leveraging this dataset, we aim to create models that are not only robust and insightful but also practical for real-world scenarios, making significant strides in advancing sleep stage detection and addressing sleep-related disorders.
4o

4.1.1 DreamT Dataset

The DREAMT dataset was selected for this study due to its comprehensive and diverse collection of physiological, demographic, and clinical data, making it highly suitable for exploring the detection of sleep stages and sleep-related disorders.

This diverse data enables advanced machine learning applications, offering opportunities for feature engineering, correlation analysis, and the development of accessible solutions for sleep monitoring using consumer-grade devices.

4.1.2 Experiments

We designed two different configurations: one consisting solely of nightly sleep data and another combining nightly sleep data with the patients' historical information. To analyze these datasets, we employed several models, considering their respective benefits and drawbacks for our data.

Random Forest
A robust ensemble learning method that constructs multiple decision trees and aggregates their predictions. It is particularly useful for handling high-dimensional data and provides feature importance metrics for interpretability. However, it may be computationally expensive for large datasets.

XGBoost (Extreme Gradient Boosting)
A highly efficient and scalable gradient boosting algorithm. It is well-suited for structured datasets and excels in predictive performance by optimizing model complexity and reducing overfitting. Its downside is that it requires careful tuning to achieve optimal results.
Deep Learning
A class of neural networks capable of capturing complex patterns and nonlinear relationships in the data. Deep learning models are highly flexible and perform well with large datasets. However, they require significant computational resources and may overfit smaller datasets without proper regularization.

Logistic Regression
A simple and interpretable linear model used for binary classification tasks. It is computationally efficient and serves as a strong baseline model. Nevertheless, its assumption of linear relationships between features and the target variable may limit its performance on complex datasets.

SGD Classifier: A linear classifier optimized using Stochastic Gradient Descent (SGD), which is efficient for large-scale datasets. It can handle both binary and multi-class classification problems. While SGD is fast and suitable for high-dimensional data, it can be sensitive to hyperparameters, especially the learning rate, and may underperform when the data has non-linear relationships.

4.1.3 Setup and Objectives

Our study involved two main experiments, each targeting specific aspects of sleep prediction and classification:

4.1.3.1 Detection of Sleep and Wake States
We evaluated the models' ability to accurately distinguish between asleep and awake conditions. This foundational task is essential for assessing the overall reliability of the system.

4.1.3.2 Estimation of Sleep Stages
We explored the models' capability to estimate the current sleep stage of a patient at a given time and under specific conditions. This experiment focused on identifying transitions between different sleep stages, providing more granular insights into sleep patterns.

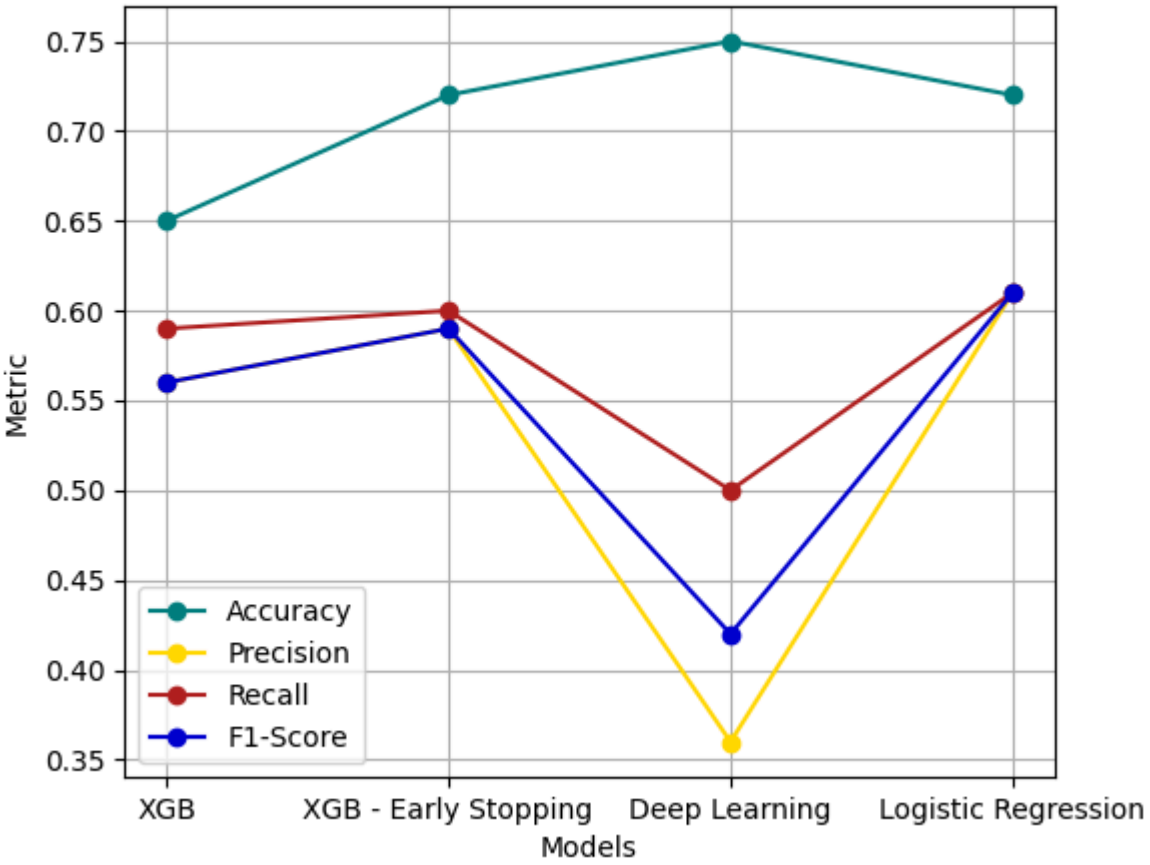
Additionally, we incorporated the **Preparation Stage** component into the experiment. For some models, we tried evaluating performance with this stage for mitigating class imbalance by incorporating data from the preparation stage, and enhancing the system's ability to model real-world sleep behaviors comprehensively.

These experiments were designed to test the models' robustness and applicability in realistic scenarios, paving the way for more accurate and accessible sleep analysis solutions.

4.2 Experiment Results

4.2.1 Detection of Sleep and Wake States

For this task, we employed three different models to evaluate their performance, including two distinct configurations of the XGB classifier to explore its potential under varying setups. The following are the results we obtained from these experiments, highlighting the strengths and limitations of each approach.



Logistic Regression proves to be the best-performing model for this dataset due to its simplicity and ability to handle the characteristics of the data effectively. Logistic Regression is particularly well-suited for datasets where the relationships between features and outcomes are relatively linear or can be captured with straightforward transformations. Its interpretability and robust handling of imbalanced classes through techniques like class weighting make it a reliable choice for this task.

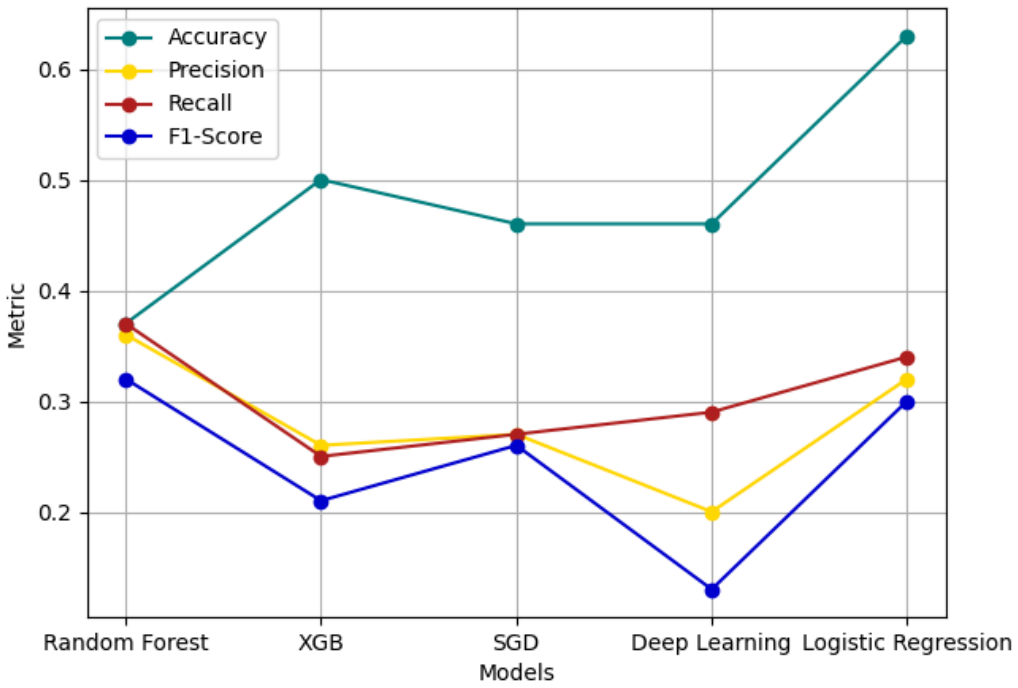
	XGB	XGB - Early Stopping	Deep Learning	Logistic Regression
Accuracy	**0.65**	**0.72**	**0.75**	**0.72**
Precision	**0.56**	**0.59**	**0.36**	**0.61**
Recall	**0.59**	**0.60**	**0.50**	**0.61**
F1-Score	**0.56**	**0.59**	**0.42**	**0.61**

On the other hand, while the Deep Learning model achieves high overall accuracy, this performance metric is misleading. The high accuracy is primarily driven by the model's tendency to predict the majority class (N2 stage) in an imbalanced dataset. This phenomenon occurs because Deep Learning models, especially when not carefully tuned for class imbalance, are prone to overfitting the majority class. The model prioritizes minimizing overall loss, often at the expense of minority classes, leading to poor precision, recall, and F1-scores for those underrepresented categories.

This limitation of Deep Learning arises from its reliance on extensive training data to generalize well and its sensitivity to class distributions. Even when class balancing techniques, such as oversampling or cost-sensitive learning, are applied, they did not fully address the issue without fine-tuning the model architecture and hyperparameters. As a result, while Deep Learning can excel in more balanced or complex datasets, its performance on this particular dataset is undermined by the dominance of the majority class, rendering it less effective compared to Logistic Regression.

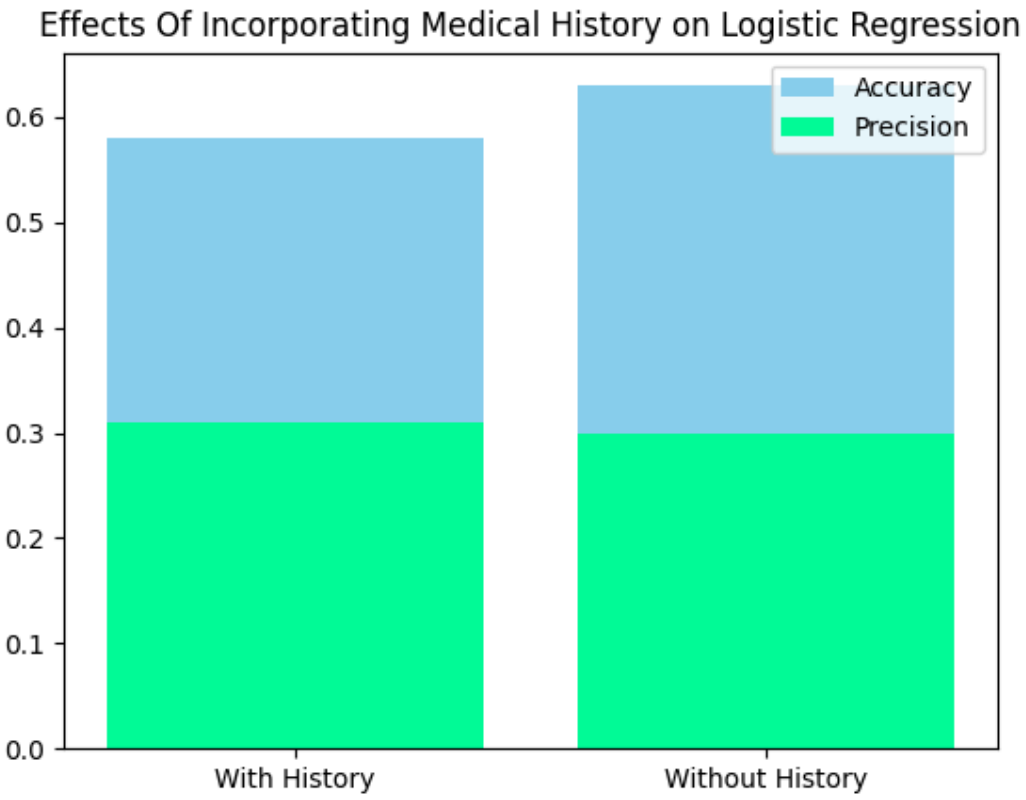
4.2.2 Sleep Stage Estimation

	Random Forest	Stochastic Gradient Descent	XGB	Deep Learning	Logistic Regression
Accuracy	0.37	0.46	0.50	0.46	0.63
Precision	0.36	0.27	0.26	0.20	0.32
Recall	0.37	0.27	0.25	0.29	0.34
F1-Score	0.32	0.26	0.21	0.13	0.30



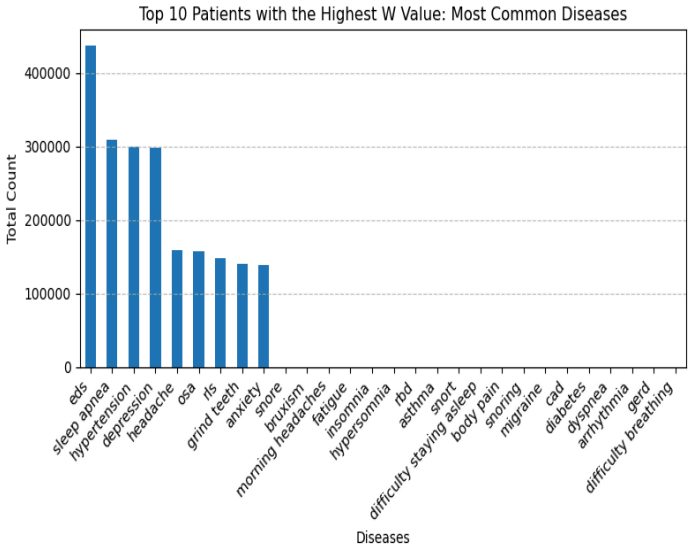
The experiment results indicate that our model struggles to predict certain classes, while performing better on others. Logistic Regression emerged as the most effective model for this task, followed by XGBoost. Although the Deep Learning model achieved relatively high accuracy, it failed to capture the underlying data relationships. The model predominantly predicts the N2 stage due to its sensitivity to class imbalance, and the class imbalance correction techniques were not sufficient to resolve this issue.

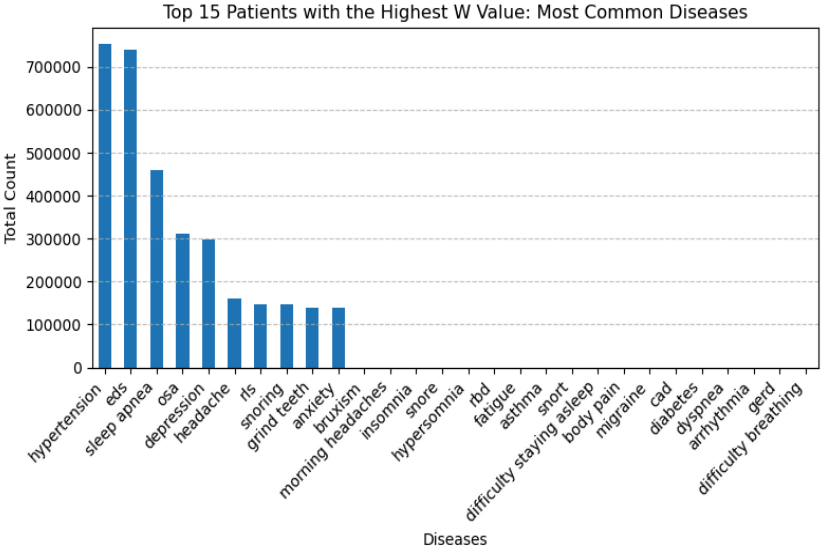
We also tested the exclusion of medical histories from the model, and surprisingly, we found that including medical history had little to no effect, or even a negative impact, on model performance. Below is the result for logistic regression:

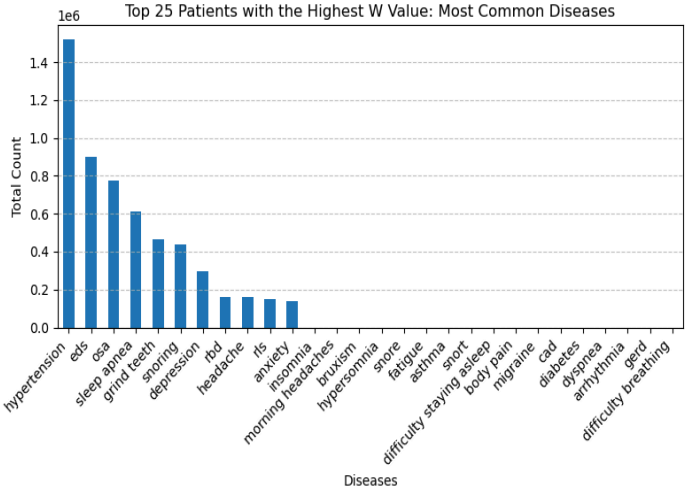


From the graph, we can infer that including patient history does not significantly improve the model's performance. The one-hot encoding of medical history negatively impacts the model's structure, leading to diminished effectiveness.

4.2.3 How is sleep apnea linked to a patient's medical history?

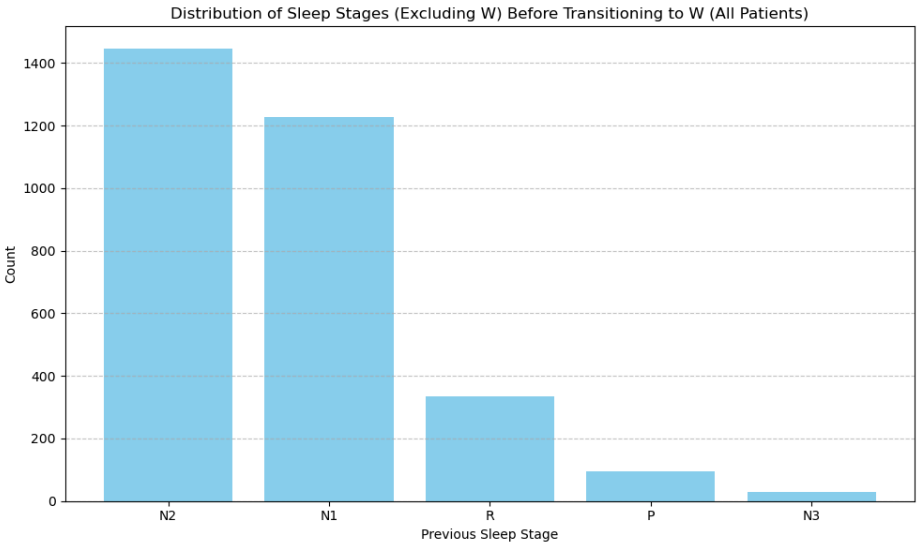






These three charts illustrate the medical history of patients with the highest occurrences of the W state, focusing on groups of the top 10, 15, and 25 patients, respectively. When looking at the top 10 patients with the most W occurrences, it is evident that sleep apnea is strongly associated with excessive daytime sleepiness (EDS). However, due to the prevalence of hypertension among patients in general, it is not possible to directly link this condition to sleep apnea in a definitive way. Overall, the data suggests that conditions like EDS, depression, bruxism, and snoring are more directly related to sleep apnea, whereas hypertension appears as a common comorbidity, likely influenced by other factors as well.

4.2.4 Analysis of the Sleep Stages Patients Transition Through Before Entering the W State



From the graph, it is evident that patients most frequently transition to the W state from the N2 sleep stage. Upon examining the dataset, we observe that half of the patients lack the N3 stage entirely. Therefore, we can infer that the transition to the W state may be primarily attributed to the inability of patients to reach the N3 stage.

5. CONCLUSION

This study successfully demonstrates the potential of leveraging physiological signals, such as electrodermal activity, heart rate variability, and accelerometer data, to detect sleep stages and analyze sleep-related disorders like obstructive sleep apnea (OSA). By utilizing the DREAMT dataset, we developed a machine learning pipeline that offers a more accessible and cost-effective alternative to traditional polysomnography (PSG). Logistic Regression emerged as the most robust and interpretable model, excelling in balancing performance metrics and handling class imbalances. Our results also underline the significance of feature engineering, with metrics like normalized heart rate and abnormal inter-beat intervals capturing essential patterns across sleep stages and apnea events. These contributions provide a foundation for advancing non-invasive sleep monitoring technologies.

However, the study also revealed some limitations. Deep learning models struggled with class imbalances, often favoring majority classes despite applying corrective techniques. Additionally, the inclusion of medical history data did not yield significant improvements in model performance, pointing to the need for more sophisticated data representations. Future work should address these challenges by refining deep learning architectures, incorporating advanced class imbalance solutions, and testing the models on more diverse and extensive datasets. Furthermore, integrating additional balanced data, particularly for underrepresented classes, could significantly enhance the predictive power and generalizability of the models. Expanding the framework to real-world applications, such as wearable devices, could further enhance the practicality of this research, enabling more inclusive and efficient sleep analysis for clinical and consumer use.

Funding

There is no funding for this publication.

Conflict of interest

The authors declare that they have no conflict of interest.

Data Availability

Tips for Data Availability Statement: https://communities.springernature.com/posts/tips-for-writing-a-dazzling-das-data-availability-statement

REFERENCES

[1] T. Giles, T. Lasserson, B. Smith, J. White, J. Wright, and C. Cates, "Continuous positive airways pressure for obstructive sleep apnoea in adults," *Cochrane Database of Systematic Reviews*, Jan. 2006, doi: https://doi.org/10.1002/14651858.cd001106.pub2.

[2] P. E. Peppard, T. Young, M. Palta, and J. Skatrud, "Prospective Study of the Association between Sleep-Disordered Breathing and Hypertension," *New England Journal of Medicine*, vol. 342, no. 19, pp. 1378–1384, May 2000, doi: https://doi.org/10.1056/nejm200005113421901.

[3] E. SHAHAR *et al.*, "Sleep-disordered Breathing and Cardiovascular Disease," *American Journal of Respiratory and Critical Care Medicine*, vol. 163, no. 1, pp. 19–25, Jan. 2001, doi: https://doi.org/10.1164/ajrccm.163.1.2001008.

[4] N. M. Punjabi *et al.*, "Sleep-Disordered Breathing and Mortality: A Prospective Cohort Study," *PLoS Medicine*, vol. 6, no. 8, p. e1000132, Aug. 2009, doi: https://doi.org/10.1371/journal.pmed.1000132.

[5] T. Young, M. Palta, J. Dempsey, J. Skatrud, S. Weber, and S. Badr, "The occurrence of sleep-disordered breathing among middle-aged adults," *The New England journal of medicine*, vol. 328, no. 17, pp. 1230–5, 1993, doi: https://doi.org/10.1056/NEJM199304293281704.

[6] Mehmet KARADAĞ and Ahmet URSAVAŞ, “Dünyada ve Türkiye’de Uyku Çalışmaları,” *Turkiye Klinikleri Archives of Lung*, vol. 8, no. 2, pp. 62–64, 2014, Accessed: Nov. 16, 2024. [Online]. Available: https://www.turkiyeklinikleri.com/article/en-dunyada-ve-turkiyede-uyku-calismalari-54777.html

[7] “How Much Does A Sleep Study Cost?,” *Sleep Foundation*, Mar. 31, 2023. https://www.sleepfoundation.org/sleep-studies/how-much-does-a-sleep-study-cost

[8] Penzel, T., Kantelhardt, J. W., Lo, C.-C., Voigt, K., & Vogelmeier, C. (2003). Dynamics of heart rate and sleep stages in normals and patients with sleep apnea. *Neuropsychopharmacology*, *28*(S1). https://doi.org/10.1038/sj.npp.1300146

[9] Wang, W. K., Yang, J., Hershkovich, L., Jeong, H., Chen, B., Singh, K., Roghanizad, A. R., Shandhi, M. M. H., Spector, A. R., & Dunn, J. (2024). Addressing wearable sleep tracking inequity: A new dataset and novel methods for a population with sleep disorders. In T. Pollard, E. Choi, P. Singhal, M. Hughes, E. Sizikova, B. Mortazavi, I. Chen, F. Wang, T. Sarker, M. McDermott, & M. Ghassemi (Eds.), *Proceedings of the fifth Conference on Health, Inference, and Learning* (Vol. 248, pp. 380–396). PMLR. https://proceedings.mlr.press/v248/wang24a.html

[10] Gaiduk, M., Perea, J. J., Seepold, R., Martinez Madrid, N., Penzel, T., Glos, M., & Ortega, J. A. (2022). Estimation of sleep stages analyzing respiratory and movement signals. IEEE Journal of Biomedical and Health Informatics, 26(2), 505–514. https://doi.org/10.1109/jbhi.2021.3099295

[11] Nochino, T., Ohno, Y., Kato, T., Taniike, M., & Okada, S. (2019). Sleep stage estimation method using a camera for home use. *Biomedical Engineering Letters, 9*(4), [pages]. https://doi.org/10.1007/s13534-019-00108-w

APPENDIX