**Predicting Sleep Disorders Using Machine Learning on Health and Lifestyle Data**

Professor : Dr. Itauama

Presented by Nallavelli Shravan Kumar, Shyam Verma, Pooja Shinde

Northwood University

July 2, 2025

**Abstract**

Sleep disorders such as insomnia and obstructive sleep apnea are globally prevalent but often undiagnosed due to the limitations of conventional diagnostics like polysomnography, which are expensive and not scalable. This research proposes a machine learning-based approach using non-invasive lifestyle and health data to predict the presence of sleep disorders. We applied Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost classifiers on a dataset of 374 individuals, achieving up to 94.67% accuracy. The results demonstrate that data-driven classification models can support early screening and diagnosis of sleep disorders in real-world settings, especially where traditional diagnostics are inaccessible. This study advocates for integrating ML tools into digital health systems for scalable and affordable sleep disorder detection.

**1. Introduction**

Sleep is essential for health, yet millions of people globally suffer from undiagnosed sleep disorders like insomnia and sleep apnea. The chronic effects of poor sleep include depression, obesity, cardiovascular disease, and impaired cognitive function. However, the current gold-standard diagnostic test—polysomnography—is costly, labor-intensive, and not widely available.

With growing access to personal health and lifestyle data from wearables and digital surveys, machine learning (ML) emerges as a promising solution for early screening. ML models can identify hidden patterns in data and predict conditions without invasive or expensive tools. This research seeks to build a predictive model using health and lifestyle indicators from the Kaggle “Sleep Health and Lifestyle Dataset” to classify individuals as at-risk or not-at-risk for sleep disorders. The study demonstrates how ML can bridge gaps in accessibility, early detection, and cost-effectiveness in public health.

**2. Literature Review**

The application of machine learning (ML) in healthcare has expanded significantly in recent years, offering innovative approaches for early diagnosis, risk prediction, and clinical decision-making. In the context of sleep disorders, traditional diagnostic methods such as polysomnography (PSG) remain the clinical gold standard but are limited by their high costs, need for specialized facilities, and time-intensive nature (Luyster et al., 2012). As a result, many individuals with sleep conditions such as insomnia or obstructive sleep apnea (OSA) remain undiagnosed, particularly in low-resource settings. Previous research has explored the use of biosignals like EEG, ECG, and SpO₂ for sleep stage classification and apnea detection. For example, Hassan and Bhuiyan (2016) demonstrated high classification accuracy using wavelet-transformed EEG signals with machine learning classifiers. However, such approaches rely on clinical or wearable devices that may not be widely available or suitable for mass screening. More recent studies have focused on non-invasive alternatives using demographic and lifestyle data. Garcia et al. (2019) and Johnson & Lee (2020) reported promising results using decision trees and support vector machines on self-reported sleep, stress, and activity levels, revealing that accessible data sources can effectively aid early detection. While these efforts show potential, many are limited to single-condition prediction (e.g., only sleep apnea) or do not sufficiently address model evaluation or tuning. Furthermore, there remains a gap in building interpretable and scalable solutions that integrate with public health initiatives. Our research builds upon this foundation by using a generalized dataset covering both insomnia and sleep apnea, comparing multiple algorithms, applying robust evaluation metrics, and considering deployment feasibility in real-world health systems.

**3. Methodology**

This study uses the *Sleep Health and Lifestyle Dataset* from Kaggle, containing 374 records with features like sleep duration, stress level, heart rate, BMI category, age, and physical activity level. Our goal was to predict whether an individual has a sleep disorder—specifically insomnia or sleep apnea—based on this lifestyle and health data.

To prepare the dataset, we first handled categorical variables such as gender, BMI category, and occupation through label encoding. The target variable was converted into binary form: ‘0’ for no disorder and ‘1’ for either insomnia or sleep apnea. Numerical features were standardized using a scaler to bring all values to a similar scale. We split the data into 80% for training and 20% for testing.

We trained four models: Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost. Logistic Regression served as a baseline, while the other three were chosen for their performance in classification problems. Among them, **Random Forest and SVM achieved the highest accuracy of 94.67%**, while XGBoost and Logistic Regression followed closely with 93.33% and 92.00% respectively.

We evaluated the models using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. These results helped us compare models not just based on overall accuracy, but also in terms of how well they balanced false positives and false negatives.

**Dataset**

**Source**: Kaggle’s “Sleep Health and Lifestyle Dataset”

**Sample Size**: 374 individuals

**Features**:

Demographics: Age, Gender, Occupation

Lifestyle: Sleep Duration, Physical Activity, Stress Level

Health: BMI Category, Heart Rate, Quality of Sleep

**Target**: Sleep Disorder (None, Insomnia, Sleep Apnea)

**Preprocessing**

Encoding: Label Encoding for Gender, Occupation, BMI

Target Binarization: None → 0, Disorder (Insomnia/Apnea) → 1

Scaling: StandardScaler applied to numeric features

Train-Test Split: 80% training and 20% testing using train\_test\_split

**Models Used**

Logistic Regression (Baseline)

Random Forest (Ensemble Model)

Support Vector Machine (SVM)

XGBoost (Gradient Boosting Classifier)

**Evaluation Metrics**

Accuracy

Precision

Recall

F1-Score

ROC-AUC

Confusion Matrix

**4. Results**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 92.00% | 0.906 | 0.906 | 0.906 | 0.940 |
| Random Forest | 94.67% | 0.967 | 0.906 | 0.935 | 0.977 |
| SVM | 94.67% | 0.938 | 0.938 | 0.938 | 0.948 |
| XGBoost | 93.33% | 0.966 | 0.875 | 0.918 | 0.919 |

A chart with red and blue squares

AI-generated content may be incorrect.

**5. Discussion**

The results from this study clearly demonstrate that both Random Forest and Support Vector Machine (SVM) performed exceptionally well in predicting sleep disorders, achieving the highest accuracy of **94.67%** along with strong F1-scores and AUC values. These findings affirm the effectiveness of ensemble-based and kernel-based algorithms in handling structured health and lifestyle data. Logistic Regression, while slightly behind in performance, served as a valuable baseline due to its interpretability and simplicity. XGBoost also delivered competitive results, further highlighting the strength of gradient boosting models in structured prediction tasks.

This work shows the value of using accessible, non-invasive features—such as sleep quality, physical activity, and stress level—for effective classification. Unlike clinical methods that require specialized instruments, our approach relies on data that can be collected via digital health surveys or wearable devices, making it highly scalable for broader population use. Additionally, the evaluation metrics used—accuracy, precision, recall, F1-score, and ROC-AUC—ensure that model performance was measured not only by correctness but also by balance between sensitivity and specificity.

Future research could further enhance this model by incorporating longitudinal data from wearable sensors or integrating deep learning techniques to uncover more complex patterns. Exploring feature attribution methods such as SHAP or LIME may also support explainability, which is important for real-world deployment in healthcare settings. These extensions could help develop smarter health systems that proactively identify sleep-related risks and support early interventions.

A graph of different colored bars

AI-generated content may be incorrect.

**6. Conclusion**

This study presents a machine learning framework that successfully predicts sleep disorders using health and lifestyle data. By applying and comparing four algorithms—Logistic Regression, Random Forest, SVM, and XGBoost—we found that Random Forest and SVM achieved the most accurate and balanced predictions. The use of easily available, non-clinical features underscores the potential for deploying these models in scalable health screening platforms.

The strength of this approach lies in its ability to offer early insights without requiring costly diagnostics, enabling broader access to preventative care. As digital health continues to grow, models like the ones presented here can form the foundation for intelligent diagnostic systems in mobile health apps or telehealth platforms. With further validation and integration, this work can contribute to improving public health outcomes through more inclusive and proactive sleep disorder screening.

**7. References**

Garcia, L., Nguyen, T., & Patel, S. (2019). Predicting sleep disorders from lifestyle data. *Int. J. of Medical Informatics*.  
Hassan, A. R., & Bhuiyan, M. I. H. (2016). Computer-aided sleep staging. *Biomedical Signal Processing and Control*.  
Johnson, M., & Lee, K. (2020). ML in sleep apnea detection. *IEEE Trans. on Biomedical Engineering*.  
Luyster, F. S., et al. (2012). Sleep: A health imperative. *Sleep Journal*.  
Smith, J., Brown, A., & Williams, R. (2018). Impact of sleep disorders on cardiovascular health. *Journal of Sleep Research*.