Advanced Data Challenge
Analyzing Sleep Data in Sleeping Beauty Syndrome
Experimental Report

Tarikwa Bedane
Rajani Jagarwal
Paris Lodron University of Salzburg
tarikwa.bedane@plus.stud.ac.at
rajani.jagarwal@plus.stud.ac.at

Abstract. Kleine–Levin Syndrome (KLS), commonly referred to as Sleeping Beauty Syndrome, is a rare neurological disorder characterized by recurrent episodes of hypersomnia accompanied by cognitive and behavioral disturbances. Due to the unpredictable and episodic nature of the disorder, early detection remains a major clinical challenge. This study investigates whether sleep and physiological data collected from wearable devices can be used to identify early warning signals preceding hypersomnia episodes. A daily time-series data set consisting of approximately one hundred days of sleep duration, sleep architecture, and heart-rate measurements from a single diagnosed KLS patient was analyzed. The research involved data cleaning, exploratory analysis, feature engineering, and the evaluation of multiple machine-learning models including Logistic Regression, Random Forest, XGBoost, and Neural Networks. The findings suggest that lag-based sleep duration and physiological features provide predictive signals. However, dataset size and subject diversity remain key limitations for clinical validation.

Keywords: Kleine–Levin Syndrome, Sleep Data, Machine Learning, Time Series Prediction, Wearable Data, Hypersomnia

1 Introduction

Kleine–Levin Syndrome (KLS), also known as "Sleeping Beauty Syndrome," is a rare neurological disorder characterized by recurrent episodes of hypersomnia. During these episodes, patients may sleep more than 15 hours per day and experience cognitive impairment, confusion, and behavioral changes [1, 4]. Between episodes, sleep patterns and daily functioning typically return to normal. The objective of this study is to analyze longitudinal sleep data collected from a diagnosed KLS patient in order to explore sleep behavior patterns, detect hypersomnia episodes, and identify meaningful indicators that may later support predictive modeling.

2 Data set Description

2.1 Data Source

The data set was collected by the research group of Prof. Manuel Schabus using the Sleep2 monitoring system [1]. Data were recorded approximately 100 consecutive days.

[1] https://sleep2.at

2.2 Data set Structure

The data set consists of daily time-series sleep measurements from a single patient with KLS. It has more than 60 features, and two recording modes are present:

- **Sensor Mode**: Automatically recorded physiological and sleep-stage data during regular nights.
- **Manual Mode**: Self-reported total sleep time, primarily during hypersomnia episodes.

Approximately 20 days correspond to suspected hypersomnia attacks and are mainly recorded in manual mode.

2.3 Feature Categories

Sleep-related variables were grouped into the following categories:

Table 1: Feature Categories

Name	Variables	Description
Sleep Duration	TST, TIB, TSP, subjective_sleep.time	Total amount of sleep obtained
Sleep Continuity	WASO, SE, sleep_interruption_count	Continuity and quality of sleep
Sleep Architecture	REM%, Deep%, Light%, NREM%, REM_SH, REM_F, DEEP_SH, DEEP_FH	Composition of different sleep stages
Circadian Timing	bed_time, sleep_time, wake_time, bed_time_deviation	Timing of sleep relative to the 24-hour cycle
Physiological Metrics	HR metrics	Physiological measures during sleep
Behavioral / Environmental Factors	sleep_quality_rating, workfree_day, workfree_day tomorrow, daily_notes	Subjective or external influences on sleep

3 Data Preprocessing

Missing data in this study were not treated purely as technical artifacts but as potentially informative signals of hypersomnia episodes.

In manual recording mode, only Total Sleep Time (TST) was available, while detailed sleep-stage variables were absent. To enable consistent analysis:

- Weekday and weekend indicators were completed to allow comparison.
- For visualization purposes, REM, Deep, and Light sleep proportions in manual entries were normalized to sum to 100%.
- Missing Sleep Score values were imputed using the dataset mean to maintain temporal continuity in plots.
- For certain features like sleep stages, sleep scores, HR bins, lag features at t-1 and t-2 days were computed and created.

4 Exploratory Data Analysis

4.1 Total Sleep Time Trends

Total Sleep Time (TST) displayed in Figure 1 pronounced spikes exceeding 15 hours during manual-mode entries. A 3-day rolling mean further highlighted sustained periods of extended sleep, consistent with hypersomnia episodes.

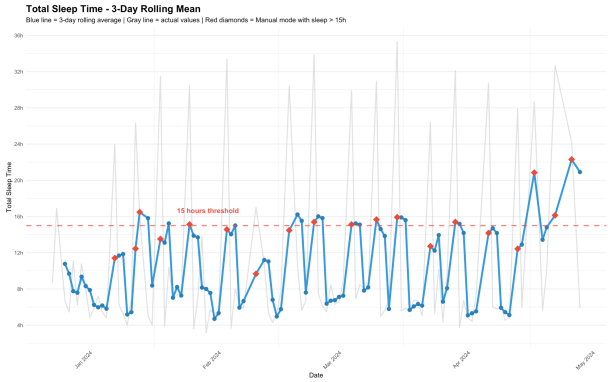
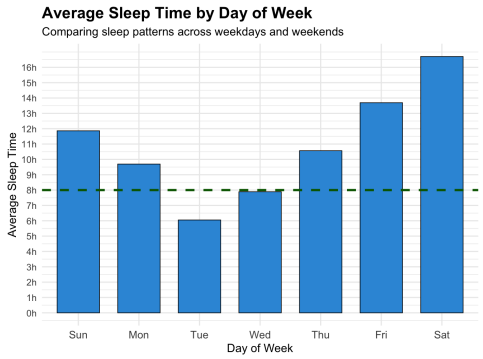


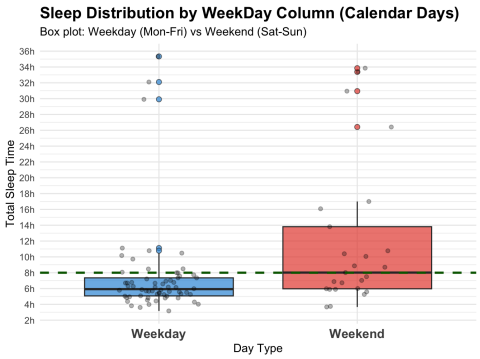
Figure 1: Sleep Stage Distribution over Time.

4.2 Weekday and Weekend Comparison

Average sleep duration differed between weekdays and weekends. Weekend sleep showed greater variability and higher median values compared to weekdays, suggesting contextual or behavioral influences on sleep duration the following two figures in Figure 2.



(a) Average Sleep Time by Weekdays



(b) Sleep Distribution by Weekdays

Figure 2: Comparison of sleep by Weekday and Weekend.

4.3 Sleep Score Analysis

Sleep Score fluctuated across the observation period with a mean value of approximately 5.04 as shown in Figure 3. Manual-mode entries exhibited noticeable deviations in score patterns compared to sensor-recorded nights, indicating altered sleep quality during hypersomnia phases [3].

4.4 Sleep Architecture Patterns

Observed sleep-stage distributions differed from typical expected values:

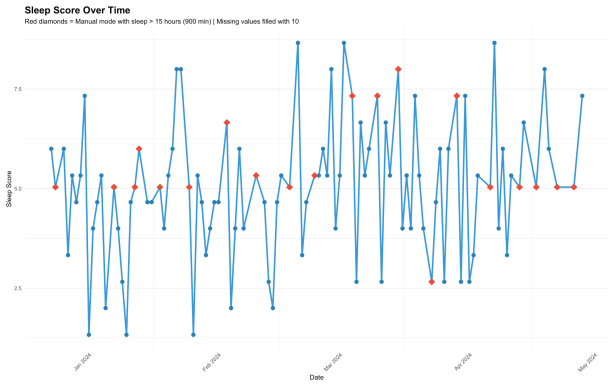


Figure 3: Sleep Score over Time.

- Expected Deep Sleep: 13–23%; Observed Mean: 12.07%.
- Expected REM Sleep: 20–25%; Observed Mean: 14.82%.
- Expected Light Sleep: 45–55%; Observed Mean: 73.10%.

The patient exhibited an elevated proportion of Light Sleep and reduced REM Sleep, suggesting disrupted sleep architecture as shown in Figure 4.

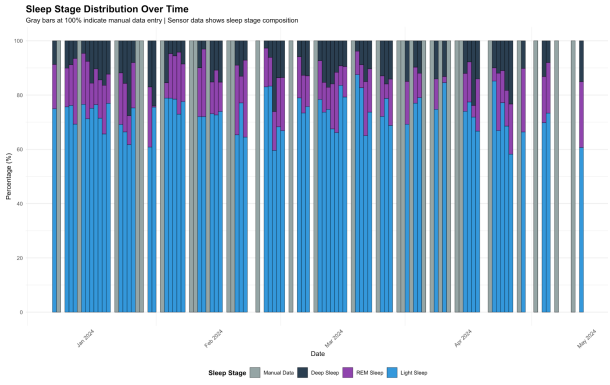


Figure 4: Sleep Stage Distribution over Time.

5 Heart Rate Data Preprocessing

Heart rate (HR) measurements were collected as pipe-separated strings containing 5-minute interval values recorded during sleep. Raw wearable data exhibited noise, missing values, and physiologically implausible spikes . Therefore, a structured preprocessing pipeline as illustrated in Figure 5 was implemented to ensure stable and meaningful physiological signals.

First, safe parsing was applied to remove invalid tokens (e.g., null, empty strings). The extracted values were converted into numerical format.

Second, a physiological filtering step was performed by restricting heart rate values to the human plausible range of 30–220 BPM. Values outside this interval were discarded. Third, artifact removal was conducted to eliminate sudden unrealistic jumps. Measurements deviating more than 40 BPM from adjacent values were considered noise and removed.

Finally, median filtering (kernel size = 3) was applied to smooth short-term fluctuations while preserving underlying trends. Figure 6 shows the comparison of original heart rate and the preprocessed heart rate data.

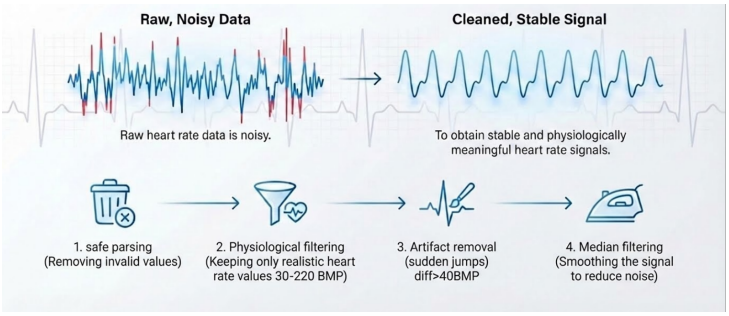
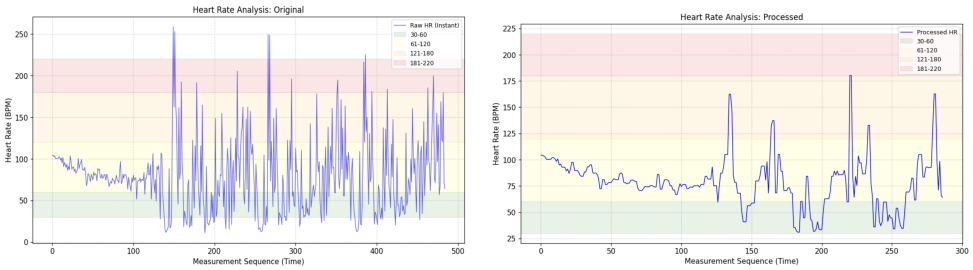


Figure 5: Heart Rate Data Preprocessing Pipeline.



(a) Heart rate Analysis Original (b) Heart rate Analysis Processed

Figure 6: Comparison of HR original and processed

After doing all the above steps, summary features were extracted from the cleaned signal, including mean heart rate, standard deviation, minimum, maximum, linear trend, percentage of missing values, and the proportion of samples within predefined heart rate bins (30–60, 60–120, 120–180, 180–220 BPM) as shown in the following Figure 7.

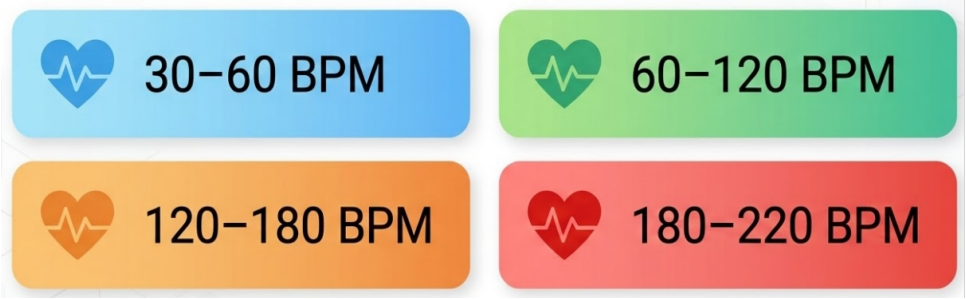


Figure 7: Heart Rate Bins.

This preprocessing pipeline ensured physiologically meaningful heart rate features suitable for downstream predictive modeling.

6 Feature Engineering and Selection

To model physiological changes preceding KLS episodes, a structured feature engineering pipeline was developed integrating sleep metrics, heart rate characteristics, temporal context, and lagged dynamics.

6.1 Sleep-Derived Features

Sleep-related variables included total sleep duration, sleep interruption counts, sleep score, and sleep stage proportions (REM%, NREM%, Deep%, Light%). Since hypersomnia is the

defining characteristic of KLS, total sleep time and derived rolling averages were considered core predictors.

6.2 Temporal Features

Temporal context was incorporated using:

- Weekday indicator
- Day of year
- Days since last attack

The variable `days_since_attack` was computed as a cumulative counter reset at the beginning of each episode, capturing cyclic tendencies observed in KLS attacks.

6.3 Lagged and Rolling Features

Because KLS episodes are not instantaneous events but evolve over multiple days, time-lagged features were constructed to incorporate short-term historical information. For selected physiological and sleep metrics, lag features at $t-1$ and $t-2$ days were created. Additionally, rolling averages over a 3-day window were computed to smooth daily fluctuations and capture short-term trends.

This temporal augmentation allowed the models to learn precursor patterns rather than relying solely on same-day measurements.

6.4 Feature Selection Technique

Given the limited dataset size and class imbalance, feature selection was performed using two complementary approaches:

1. **Random Forest Feature Importance:** Tree-based models were used to estimate the relative contribution of each feature. Features with consistently high importance scores were retained for modeling.
2. **Correlation Analysis:** Pearson correlation coefficients between candidate features and the future attack label were examined to identify variables showing early predictive signal.

These strategies ensured that both nonlinear interactions (captured by tree-based models) and direct linear relationships were considered during feature refinement.

6.5 Final Selected Features

Based on Random Forest feature importance and correlation analysis with the next-day attack label, a refined subset of predictors was retained for modeling.

The most consistently important features included total sleep duration and its lagged values (particularly `stat_total_sleep_time_minutes_lag1` and `lag2`), sleep score, selected heart rate bin percentages (e.g., `hr_bin_30_60_pct`, `hr_bin_120_180_pct`), and sleep-stage composition metrics such as `NREMPercent`.

The duration of sleep and its temporal delays showed the strongest predictive signal, confirming the clinical relevance of hypersomnia in KLS. Heart rate distribution and sleep architecture features provided complementary physiological insight, suggesting autonomic and structural sleep changes precede episode onset.These core features formed the basis of the final modeling framework.

7 Problem Formulation and Time-Series Cross Validation

7.1 Prediction Objective

The objective of this study is to predict whether a KLS episode will occur on the following day. This formulation transforms the problem into a binary next-day classification task.

Let is_attack_t denote whether day t corresponds to a hypersomnia episode. The prediction target was defined as:

$$target_t = is\_attack_{t+1}$$

Thus, the model learns from historical physiological and sleep features observed at time t to predict the probability of an attack occurring at time $t + 1$.

This formulation enables early detection of prodromal physiological changes preceding KLS episodes.

7.2 Time-Series Cross-Validation

Because the dataset consists of chronologically ordered daily observations, classical random train–test splitting would introduce temporal leakage by allowing future information to influence past predictions. To avoid this issue, a time-series-based cross-validation strategy was adopted [2].

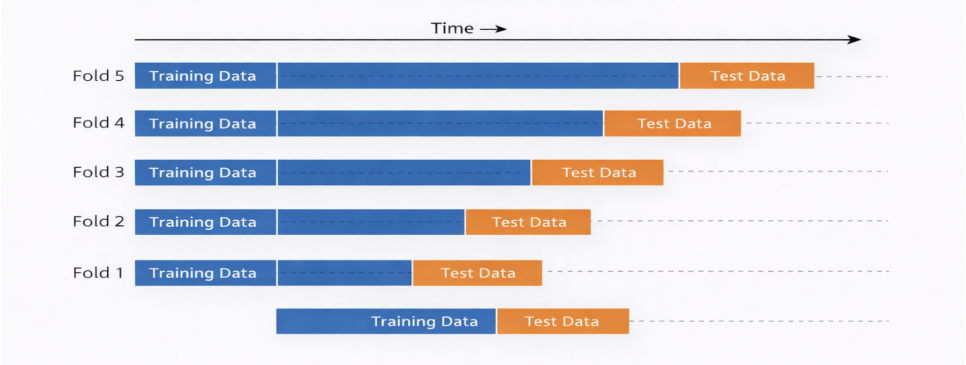


Figure 8: Time-Series Cross-Validation Folds.

An expanding window (rolling-origin) validation scheme, shown in Figure 8, was used. In this setup, the training window begins with the earliest observations and progressively expands over time, while the testing window always consists of the immediately subsequent unseen segments [2, 6].

Formally, for fold k:

- Training set: days 1 to t_k
- Testing set: days $t_k + 1$ to $t_k + h$

where h denotes the testing horizon.

This procedure was repeated across multiple folds, ensuring that each test segment chronologically follows its corresponding training segment. No shuffling was applied at any stage.

This approach mimics real-world forecasting conditions, where models are trained only on past information and evaluated on future unseen data. It also enables assessment of model stability across different temporal segments of the dataset.

7.3 Model Performance Evaluation Metrics

KLS episodes are rare events. In the processed dataset, only 20 out of 126 days were labeled as attack days, resulting in a significantly imbalanced class distribution.

Because of this imbalance, overall accuracy alone is not an appropriate performance metric. Instead, emphasis was placed on metrics that reflect the model's ability to correctly detect rare attack events.

The following evaluation metrics were used:

- **Recall (Sensitivity):** Ability to correctly detect attack days.
- **Precision:** Reliability of predicted attack days.
- **F1-score:** Harmonic mean of precision and recall.
- **PR-AUC:** Area under the Precision–Recall curve, suitable for imbalanced data sets [5].

These metrics provide a more clinically meaningful assessment of predictive performance in rare event detection scenarios.

8 Machine Learning Models and Evaluation

8.1 Models Evaluated

Four supervised learning models were evaluated:

Logistic Regression Logistic regression was used as a baseline linear classifier. To address class imbalance, balanced class weighting was applied. The model was trained with 1000 maximum iterations to ensure convergence.

Random Forest Random Forest was used to capture nonlinear interactions between sleep metrics and physiological features. The model consisted of 300 decision trees with a maximum depth of 8 and balanced class weighting to handle rare attack events.

XGBoost (Gradient Boosting) XGBoost was used to model complex feature interactions through gradient-boosted decision trees. The model was configured with depth-limited trees (max_depth=8), a learning rate of 0.01, and imbalance-aware weighting via scale_pos_weight.

Neural Network A feedforward neural network with 2 hidden layer was implemented to capture nonlinear relationships in physiological signals. The architecture consisted of:

- Input layer (engineered features)
- Hidden layers of size 64 and 32 with ReLU activation
- Dropout regularization (0.3)
- Sigmoid output layer for binary classification
- Adam optimizer (learning rate = 0.001), loss function of binary cross-entropy

The model was trained for 30 epochs.

8.2 Deep Learning Models

Recurrent architectures including RNN and LSTM were also tested to learn sequential temporal dependencies. However, due to limited dataset size and the rarity of KLS episodes, these models exhibited unstable and unreliable performance. Consequently, classical machine learning models were prioritized.

9 Model Performance Results

9.1 Logistic Regression

Logistic Regression demonstrated the most stable performance across temporal folds. With balanced class weighting, the model consistently achieved high recall while maintaining competitive precision.

Results are shown in the Table 2.

Table 2: Results with each fold - Logistic Regression

Fold	**Accuracy**	**Precision**	**Recall**	**F1-score**	**PR-AUC**	**Confusion-Matrix**
1	0.80	0.00	0.00	0.00	0.53	[[16, 1], [3, 0]]
2	0.80	0.42	1.00	0.60	0.90	[[13, 4], [0, 3]]
3	0.90	1.00	0.33	0.50	0.86	[[17, 0], [2, 1]]
4	0.95	0.75	1.00	0.86	0.98	[[16, 1], [0, 3]]
5	0.95	0.80	1.00	0.89	1.00	[[15, 1], [0, 4]]

The most influential features identified by the model included:

- `sleep_score_averaged`
- `stat_total_sleep_time_minutes_lag1`
- `stat_total_sleep_time_minutes_lag2`
- `hr_bin_120_180_pct`
- `stat_sleep_interruption_minutes`

These features indicate that both recent sleep duration patterns and heart rate intensity distribution contribute to predicting upcoming KLS episodes.

9.2 Random Forest

Random Forest achieved perfect classification in later folds, though earlier folds showed occasional missed attack cases, resulting in lower stability compared to Logistic Regression.

Results are shown in the Table 3.

The most important features identified by Random Forest were:

- `stat_total_sleep_time_minutes`
- `sleep_score_averaged`
- `stat_bed_time_deviation_minutes`
- `sleep_score`

Table 3: Results with each fold - Random Forest

Fold	Accuracy	Precision	Recall	F1-score	PR-AUC	Confusion-Matrix
1	0.85	0.00	0.00	0.00	0.46	[[17, 0], [3, 0]]
2	0.90	1.00	0.33	0.50	0.92	[[17, 0], [2, 1]]
3	0.85	0.00	0.00	0.00	0.80	[[17, 0], [3, 0]]
4	0.90	1.00	0.33	0.50	1.00	[[17, 0], [2, 1]]
5	1.00	1.00	1.00	1.00	1.00	[[16, 0], [0, 4]]

- `stat_total_sleep_time_minutes_lag2`

This suggests that both absolute sleep duration and circadian deviation measures are strong nonlinear predictors of attack onset.

9.3 XGBoost

XGBoost demonstrated similar behavior to Random Forest, achieving perfect metrics in later folds but showing variability in earlier training windows. Results are shown in the Table 4.

Table 4: Results with each fold - XGBoost

Fold	Accuracy	Precision	Recall	F1-score	PR-AUC	Confusion-Matrix
1	0.75	0.00	0.00	0.00	0.45	[[15, 2], [3, 0]]
2	0.90	0.67	0.67	0.67	0.98	[[16, 1], [1, 2]]
3	0.85	0.00	0.00	0.00	0.67	[[17, 0], [3, 0]]
4	0.90	1.00	0.33	0.50	1.00	[[17, 0], [2, 1]]
5	1.00	1.00	1.00	1.00	1.00	[[16, 0], [0, 4]]

The most influential features for XGBoost included:

- `sleep_score`
- `NREMPercent`
- `hr_bin_30_60_pct`
- `stat_total_sleep_time_minutes`
- `sleep_interruption_count`

These results indicate that both sleep architecture (NREM percentage) and heart rate distribution play important roles in capturing pre-attack physiological changes.

9.4 Neural Network

The neural network showed lower and less stable performance across folds. Important contributing features included sleep duration metrics and heart rate bin features, though the model exhibited difficulty generalizing due to limited sample size.

Results are shown in the Table 5.

Table 5: Results with each fold - XGBoost

Fold	Accuracy	Precision	Recall	F1-score	PR-AUC	Confusion-Matrix
1	0.85	0.00	0.00	0.00	0.15	[[17, 0], [3, 0]]
2	0.75	0.33	0.67	0.44	0.27	[[13, 4], [1, 2]]
3	0.85	0.00	0.00	0.00	0.15	[[17, 0], [3, 0]]
4	0.95	1.00	0.67	0.80	0.72	[[17, 0], [1, 2]]
5	0.95	1.00	0.75	0.86	0.80	[[16, 0], [1, 3]]

9.5 Overall Model Comparison

Although tree-based models achieved perfect results in the final fold, Logistic Regression demonstrated the most consistent performance across temporal folds. Given the small dataset and rare-event nature of KLS, robustness and stable recall are more clinically meaningful than isolated peak performance. Across all models, recurrent patterns emerged: total sleep duration, sleep score, lagged sleep features, and heart rate bin distributions were consistently ranked as important predictors. This suggests that short-term sleep dynamics and autonomic regulation changes precede KLS episodes based on our data set.

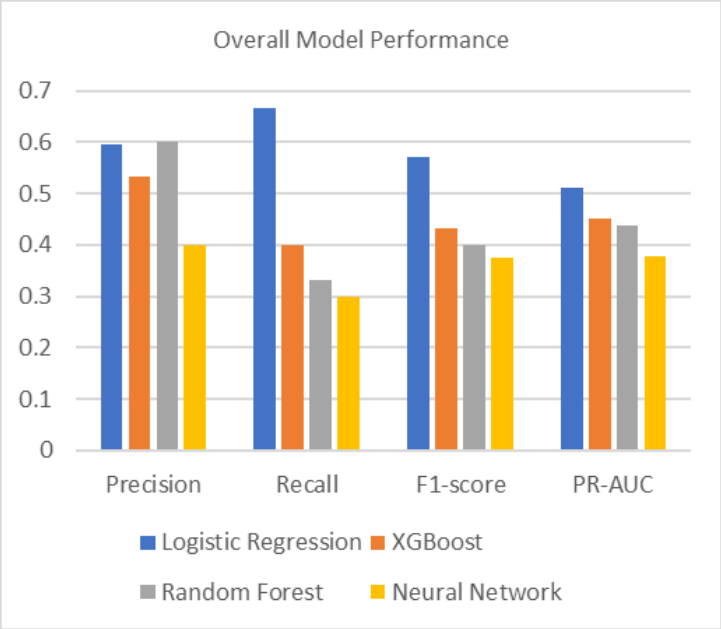


Figure 9: Overall Model Performance.

10 Limitations

Despite promising results, several limitations must be acknowledged. The final processed dataset contained 126 daily observations, of which only 20 corresponded to KLS attack days. Such a small sample size limits statistical power and increases the risk of overfitting, particularly for more complex models.

All data were collected from a single diagnosed KLS patient. Consequently, the findings may not generalize to other individuals with KLS, whose physiological patterns and attack dynamics may differ.

KLS episodes are rare events, resulting in a strongly imbalanced class distribution. Although class weighting and appropriate evaluation metrics were applied, imbalance may still influence model learning dynamics and stability.

Attack labeling was based primarily on prolonged sleep duration (greater than 900 minutes). While hypersomnia is a defining symptom of KLS, episode onset may involve additional clinical features not fully captured in wearable data.

Finally, heart rate and sleep-stage data were obtained from wearable devices, which are subject to measurement noise, missing values, and sensor artifacts. Although preprocessing steps were applied, residual measurement uncertainty may remain.

11 Conclusion

This study investigated whether sleep-derived and physiological features can be used to predict next-day KLS episodes. The results demonstrate that sleep duration, sleep score, sleep-stage composition (particularly NREM and REM percentages), and heart rate distribution metrics can be integrated to model physiological changes associated with KLS episodes.

Incorporating time-lagged features (1–2 days) significantly improved predictive performance, indicating that KLS attacks are preceded by short-term physiological shifts rather than occurring abruptly. This supports the hypothesis that prodromal patterns exist and can be detected using machine learning techniques.

Across temporal folds, Logistic Regression demonstrated the most stable and reliable performance, while tree-based ensemble models achieved strong results in later folds. These findings confirm that machine learning methods are capable of identifying patterns preceding KLS attacks, although model robustness varies.

Despite encouraging results, reliable clinical prediction requires larger multi-patient datasets, more precise episode labeling, and extended longitudinal monitoring. Future work should focus on expanding the dataset, incorporating additional physiological markers, and validating predictive models in real-world clinical settings.

References

[1] Arnulf, I., Lin, L., Gadoth, N., et al.: Diagnosis, disease course, and management of patients with kleine–levin syndrome. Brain **135**(4), 1090–1103 (2012)

[2] Bergmeir, C., Benítez, J.M.: A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis **120**, 70–83 (2018)

[3] Carskadon, M.A., Dement, W.C.: Normal human sleep: An overview. In: Principles and Practice of Sleep Medicine. Elsevier, 5 edn. (2011)

[4] Ramdurg, R.: Kleine–levin syndrome: Etiology, diagnosis, and treatment. Journal of Clinical Neuroscience **17**(8), 957–961 (2010)

[5] Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE **10**(3) (2015)

[6] Tsay, R.S.: Analysis of Financial Time Series. Wiley, 3 edn. (2010)