Indonesian Journal of Electrical Engineering and Computer Science
Vol. 44, No. 1, October 2026, pp. 49∼62
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v44.i1.pp49-62

An explainable machine learning framework for personalized sleep health assessment and behavioral recommendations

Neloy Pramanik Supto
Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

Article Info

Article history:

Received Apr 26, 2026
Revised Aug 20, 2026
Accepted Sep 10, 2026

Keywords:

Behavioral modeling
CatBoost
Explainable machine learning
Personalized recommendation
SHAP
Sleep health index
Sleep regularity

ABSTRACT

Traditional sleep research emphasizes sleep duration or clinical disorder detection, often overlooking the behavioral and temporal factors that shape everyday sleep health. This study proposes an explainable and personalized machine learning (ML) framework that models sleep health as a multidimensional behavioral construct through a continuous, engineered sleep health index (SHI) integrating timing, variability, and routine consistency. A large-scale behavioral sleep dataset ($n = 52{,}610$) was transformed into interpretable indicators including sleep regularity, bedtime variability, routine consistency, and late-night habit intensity. To control target leakage, the five SHI component variables and the raw terms that directly define them were excluded from the predictor set, and all models were evaluated under this reduced predictor configuration. Five ML models were comparatively evaluated using error-based metrics with five-fold cross-validation. CatBoost achieved the strongest held-out performance ($R^2 = 0.951$, RMSE $= 0.013$, MAE $= 0.011$), substantially below the near-perfect fit obtained when SHI components are included directly. A hybrid feature-selection strategy combining mutual information, Boruta, and SHapley Additive exPlanations (SHAP) was applied, and explainability was addressed through global and local SHAP analyses. Global explanations consistently identified sleep debt, late-night habit intensity, and social jetlag as the most influential predictors, with the ranking consistent across resamples. An autoencoder–K-means procedure identified two interpretable behavioral profiles. Because the SHI is an engineered composite derived from related behavioral variables, these results represent prediction of the proposed index rather than independent clinical validation of sleep health; the framework is offered as an exploratory decision-support tool pending external and longitudinal validation.

This is an open access article under the CC BY-SA license.



Corresponding Author:

Neloy Pramanik Supto
Department of Computer Science and Engineering, Daffodil International University
Mirpur-1207, Dhaka, Bangladesh
Email: pramanik15-4349@diu.edu.bd

1. INTRODUCTION

Sleep health is one of the most crucial determinants of mental, physical, and cognitive well-being. Poor sleep quality and irregular sleep–wake behavior are strongly associated with stress, anxiety, and metabolic disorders [1], [2]. Despite increasing awareness, global reports indicate that modern work patterns, digital exposure, and inconsistent routines have led to a significant rise in behavioral sleep problems [3].

Traditional sleep research has primarily focused on physiological signals such as electroencephalography (EEG) or actigraphy to detect disorders [4], [5], but these approaches often overlook the broader behavioral and temporal context that shapes sleep health. Understanding sleep as a behavioral system—where timing, regularity, and habits collectively influence overall sleep quality—requires a data-driven yet interpretable framework capable of both prediction and explanation.

To address these limitations, this study proposes an explainable, behavior-aware, and personalized sleep health assessment framework. Rather than treating sleep health as a binary or duration-centric outcome, the proposed approach models sleep as a multidimensional behavioral system that integrates timing, regularity, variability, routine consistency, and habit-related factors into a continuous sleep health index (SHI). By combining robust feature engineering, interpretable machine learning (ML), and individual-level explanation, this framework aims to support transparent, data-driven, and actionable sleep health assessment suitable for personalized decision support.

Related work and literature gap. ML techniques have been widely applied in sleep research, mainly in sleep-stage classification, disorder detection, and physiological signal analysis [6]. Earlier studies have shown that classical ML models such as support vector machines, k-nearest neighbors, and random forests, as well as deep learning architectures such as convolutional and long short-term memory networks, can provide strong predictive performance using EEG, heart-rate variability, actigraphy, and wearable sensor data for classification, detection, and preliminary screening of sleep disorders [4], [7]-[15]. More recent work has extended these efforts toward metaheuristic feature optimization [16], [17] and explainable prediction of sleep-related outcomes in specific populations [18], [19]. Nevertheless, several significant limitations remain in the current ML-based literature on sleep.

First, a substantial portion of research emphasizes accuracy and reports results in a "black-box" manner, prioritizing classification or prediction performance while offering little interpretability. Although deep learning models often achieve high accuracy, they seldom provide clear explanations of how behavioral or temporal factors relate to sleep outcomes, which limits their usability and credibility in clinical settings [4], [8], [15].

Second, sleep assessment in prior work commonly focuses on restricted metrics such as sleep duration, sleep efficiency, and physiological stage transitions. While these measures are practically important, they fail to capture the full behavioral picture of sleep health, which also includes bedtime variability, weekday–weekend misalignment, routine consistency, and late-night lifestyle habits. The importance of sleep timing and regularity for cognitive and metabolic health has been highlighted in systematic reviews [1], [20], yet these factors are rarely incorporated into unified predictive frameworks. Sleep is often portrayed as an isolated physiological outcome rather than a multidimensional system of behaviors.

Third, personalization and actionable guidance remain underdeveloped. Most existing studies focus on population-level modeling and report aggregated performance metrics, offering limited support for individualized interpretation or behavior-specific guidance. Although subject-specific or personalized models have improved accuracy in some contexts [21], [22], they rarely combine explainability, behavioral interpretation, and recommendation generation within a single framework. Consequently, the translation of ML-based sleep models into practical decision-support tools is still limited [3].

In summary, despite considerable progress in predictive sleep modeling, an explainable, behavior-aware, and personalized approach to sleep health assessment is still lacking. Few studies provide a continuous, interpretable SHI that integrates multidimensional behavioral indicators with individual-level explanations and actionable guidance. Bridging this gap requires an approach that is methodologically rigorous yet interpretable and application-oriented—the aim of the present study.

Objectives and contributions. The primary objective of this research is to design an explainable and personalized ML framework that quantifies and interprets sleep health using behavioral and temporal data. Specifically, this study aims to: (i) construct a composite SHI integrating sleep efficiency, regularity, duration sufficiency, routine consistency, and late-night habit burden; (ii) develop interpretable ML models to predict the SHI using explainability-driven feature selection based on mutual information, Boruta, and SHapley Additive exPlanations (SHAP); (iii) provide individual-level insight into behaviors associated with higher or lower predicted sleep health; (iv) identify data-driven behavioral profiles through autoencoder-based clustering to support group-level understanding; and (v) generate rule-based behavioral guidance to support sleep-hygiene decision-making.

The contribution is primarily framework-level integration and the formulation of a behavior-oriented composite sleep-health metric, rather than the introduction of a novel learning algorithm. A central methodological element, distinguishing this work from the original design, is explicit leakage control: the five SHI component variables and the raw terms that directly define them are excluded from the predictor set, so that reported performance is not obtained by deterministic reconstruction of the index. As discussed in section 3.2, several retained behavioral variables nevertheless remain conceptually or mathematically related to SHI components, and the resulting task is therefore interpreted as prediction of an engineered composite index from related behavioral information rather than as independent clinical outcome prediction. The statement of significance is summarized in Table 1.

Table 1. Statement of significance

Aspect	Summary
Problem or issue	Sleep-health research has emphasized physiological signal analysis and disorder detection, often neglecting behavioral regularity and lifestyle-driven variability. Existing models act as "black boxes," offering high accuracy but limited interpretability.
What is already known	Prior ML studies classify sleep stages and detect disorders using EEG or wearable data but rarely model behavioral or temporal dimensions such as bedtime variability, routine stability, or late-night habits.
What this paper adds	An explainable, behavior-aware framework that quantifies sleep health through a continuous engineered SHI integrating timing, regularity, and habit factors, using interpretable ML (SHAP + CatBoost) with an explicit leakage-control analysis, enabling both population-level interpretation and personalized decision-support guidance.
Who would benefit	Researchers, digital-health developers, and behavioral scientists seeking interpretable, data-driven, and personalized tools for exploratory assessment of sleep behavior.

2. METHOD

The overall framework transforms raw survey responses into interpretable behavioral features, constructs a continuous SHI, and applies explainable ML for prediction, personalized explanation, and behavioral clustering. The complete pipeline is illustrated in Figure 1.

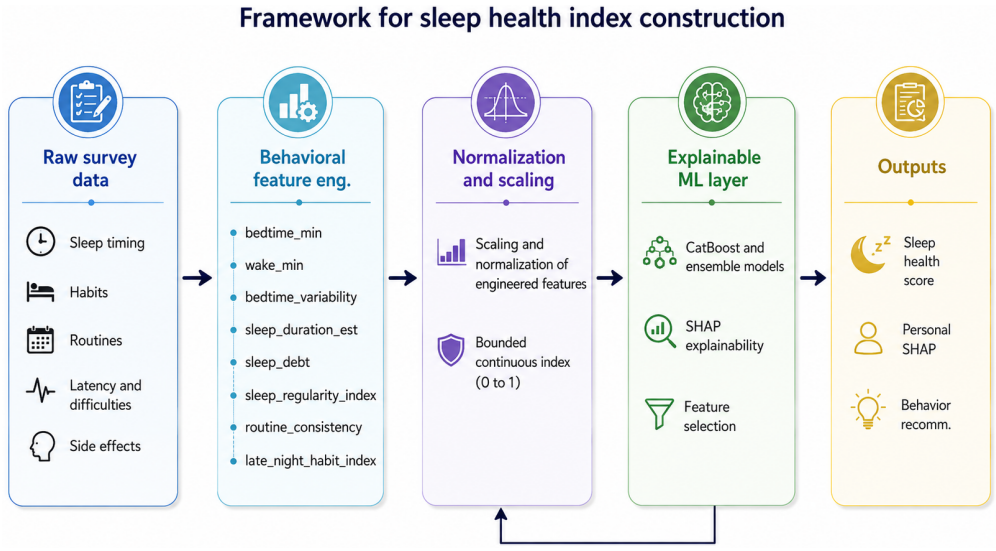

Figure 1. Modeling framework for SHI construction

2.1. Dataset description

This study uses a large-scale survey-based dataset capturing habitual sleep behaviors, daily routines, and lifestyle-related factors. The dataset comprises 52,610 individual responses, each representing a single participant. Data were collected anonymously through online questionnaires administered via Google Forms and in-person surveys across several districts of Bangladesh, with no personally identifiable information recorded.

Participation was voluntary and recruitment followed a convenience-sampling approach; no eligibility screening beyond voluntary consent was applied. Because the collection mode was not recorded on a per-response basis, the exact proportion of online versus in-person responses and the demographic representativeness of the sample relative to the national population could not be established, and this is treated as a limitation in section 3.11. The full survey instrument is available in [23].

The dataset includes 20 variables covering demographic characteristics, anthropometric attributes, weekday and weekend sleep timing, sleep duration, sleep onset latency, sleep-related difficulties, behavioral habits, coping strategies, and subjective sleep-environment comfort. An overview of the variables and their analytical roles is presented in Table 2. Most variables are categorical or ordinal, reflecting real-world questionnaire design, while a limited subset is continuous. Data-quality assessment indicated minimal missing values, primarily in the anthropometric fields (one missing weight and six missing height entries). Missing numerical values were imputed using the median, while missing categorical responses were retained as a separate category. Outliers in height and weight were identified using the interquartile-range method and capped to biologically plausible limits. Sleep-timing responses, originally reported as time intervals, were converted into minute-based numerical representations to enable quantitative estimation of sleep duration, regularity, and variability.

Table 2. Survey variables and their analytical roles

Survey variable	Data type	Behavioral domain	Analytical role
Age	Categorical	Demographic	Control variable
Gender	Categorical	Demographic	Control variable
Occupation	Categorical	Socio-behavioral	Contextual factor
Height	Continuous	Anthropometric	BMI computation
Weight	Continuous	Anthropometric	BMI computation
Weekday bedtime	Categorical (range)	Sleep timing	Converted to minutes
Weekday wake-up time	Categorical (range)	Sleep timing	Converted to minutes
Weekend bedtime	Categorical (range)	Sleep timing	Converted to minutes
Weekend wake-up time	Categorical (range)	Sleep timing	Converted to minutes
Average sleep duration	Categorical	Sleep quantity	Duration estimation
Sleep onset latency	Categorical	Sleep efficiency	Latency estimation
Difficulty falling asleep	Categorical	Sleep difficulty	Behavioral indicator
Breathing difficulty	Categorical	Sleep health	Health-related factor
Restless movements	Categorical	Sleep disturbance	Health-related factor
Medical condition	Categorical	Sleep health	Contextual factor
Reasons for sleeping late	Multi-label	Sleep habit	Habit intensity score
Coping strategies	Multi-label	Behavioral adaptation	Coping score
Side effects of poor sleep	Multi-label	Sleep impact	Severity score
Concentration difficulty	Categorical	Daytime impact	Outcome-related signal
Sleep environment comfort	Ordinal	Environmental factor	Comfort normalization

2.2. Feature engineering

Raw survey responses were transformed into interpretable behavioral indicators. Sleep-timing ranges were mapped to midpoint minutes on a 24-hour scale, with post-midnight times offset by 1440 minutes to preserve ordering. Given bedtime b and wake time w (in minutes), sleep duration D is computed as

$$D = \begin{cases} (1440 - b) + w, & \text{if } w < b \\ w - b, & \text{otherwise} \end{cases} \tag{1}$$

Weekday–weekend variability is defined for bedtime, wake time, and duration as $V_b = |b_{wd} - b_{we}|$, $V_w = |w_{wd} - w_{we}|$, and $V_D = |D_{wd} - D_{we}|$. Social jetlag is computed as the absolute difference in mid-sleep points, where midsleep $= (b + D/2) \bmod 1440$. Sleep debt is estimated relative to a 7.5-hour (450-minute) reference:

$$\text{debt} = \max(0,\ 450 - D) \tag{2}$$

The sleep regularity index (SRI) captures cross-day timing stability and is normalized to $[0, 1]$:

$$\text{SRI} = 1 - \frac{V_b + V_w}{2880} \tag{3}$$

Sleep onset latency was mapped to representative minutes (10, 23, 45, 75), and sleep efficiency was estimated as the ratio of effective sleep to total time in bed:

$$\text{efficiency} = \frac{D}{D + \text{latency}} \tag{4}$$

Routine consistency aggregates timing and duration variability into a single normalized measure:

$$\text{routine} = 1 - \frac{V_b + V_w + V_D}{4000} \tag{5}$$

Duration sufficiency measures attainment of the recommended duration, sufficiency $= \min(1, D/450)$. Multi-label responses on late-night reasons, coping strategies, and side effects were transformed into weighted additive scores and min–max normalized. The late-night habit index (LNHI) aggregates these normalized scores:

$$\text{LNHI} = 0.5\,\widehat{s}_{\text{delay}} + 0.2\,\widehat{s}_{\text{coping}} + 0.3\,\widehat{s}_{\text{side}} \tag{6}$$

where hats denote normalized scores. Body mass index was computed from weight and height and clipped to $[12, 60]$, and subjective comfort ratings were normalized to $[0, 1]$. These definitions establish the mathematical construction of the engineered target and also clarify which predictors are directly or indirectly related to its components, a distinction examined explicitly in section 2.4.

2.3. Sleep health index construction

To enable a continuous assessment of sleep behavior, a composite SHI was defined as the primary modeling target. The SHI is an engineered composite sleep-health metric—not a clinically standardized or externally validated instrument—and integrates five normalized components: sleep efficiency E, sleep regularity R, routine consistency C, duration sufficiency S, and late-night habit burden H, according to

$$\text{SHI} = w_1 E + w_2 R + w_3 C + w_4 S - w_5 H, \qquad \text{SHI} \in [0, 1] \tag{7}$$

The weights are informed by sleep-science evidence rather than assigned arbitrarily. Because sleep regularity has been shown to be associated with health outcomes more strongly than duration alone [1], [20], [24], regularity and efficiency receive the highest weights ($w = \{0.25, 0.30, 0.15, 0.20, 0.10\}$ for E, R, C, S, H). To examine dependence on the specific weighting, three alternative schemes (equal, efficiency-heavy, and regularity-heavy) were evaluated; the resulting indices correlate strongly with the primary SHI (Pearson $r > 0.97$ in all cases). This high correlation indicates that the relative ordering of participants is robust to the tested weighting assumptions. The analysis therefore supports weight-sensitivity robustness, but does not constitute external or psychometric validation of the SHI. Higher SHI values indicate more regular and sufficient self-reported sleep behavior as captured by the index.

2.4. Predictor configuration and leakage control

A key methodological safeguard concerns the relationship between the SHI and the predictor set. Because the SHI is a deterministic function of a closed set of normalized components, using those same components as predictors would make the task a reconstruction problem and inflate performance. To avoid this, the five SHI components and the raw terms that directly define them (sleep efficiency, sleep regularity, routine consistency, duration sufficiency, the late-night habit index, and the timing-variability and weekday-duration terms entering them) were excluded from the predictor set. The models were instead trained on 19 behavioral, temporal, symptom, and contextual features—raw bedtime and wake times, weekend duration, social jetlag, mid-sleep points, sleep debt, delay intensity, coping and side-effect scores, symptom indicators, BMI, environment comfort, and medical-condition status.

It is important to state explicitly that this reduced predictor configuration does not render the predictor set statistically independent of the SHI. Several retained variables remain mathematically or conceptually related to SHI components. Sleep debt, $\max(0, 450 - D)$, and duration sufficiency, $\min(1, D/450)$, are both functions of sleep duration D; likewise, the retained delay-intensity, coping, and side-effect scores are the same normalized quantities that are aggregated into the LNHI. Consequently, the study excludes the SHI component variables themselves from the predictor set and evaluates performance under this reduced predictor configuration, and the resulting task should be interpreted as prediction of an engineered composite index from related behavioral information rather than as independent clinical outcome prediction. Section 3.2 reports the corresponding empirical comparison.

2.5. Feature selection

A multi-stage feature-selection strategy combining statistical relevance, model-based importance, and explainability was applied entirely on the training partition to avoid selection leakage. First, mutual information ranked features by their non-linear dependency with the SHI. Second, Boruta selection within a random-forest framework retained features demonstrating statistically significant relevance against randomized shadow features. Finally, SHAP analysis assessed both the magnitude and direction of feature contributions to model output. Features confirmed by Boruta and highly ranked by mutual information and SHAP were retained as the stable, interpretable subset. The complete feature-selection procedure was carried out on the 80% training partition, and the held-out test set remained untouched until final evaluation.

2.6. Machine learning models

To model the SHI, five algorithms were comparatively evaluated. Linear regression served as a baseline to establish a lower bound and, importantly, to test whether the target could be reconstructed linearly from the predictor set. Random forest was included for its robustness to noise and ability to model feature interactions. XGBoost and LightGBM were included as gradient-boosting techniques suited to structured behavioral data. CatBoost, whose ordered boosting reduces overfitting and prediction bias, was evaluated as the primary candidate. Hyperparameters were tuned to balance accuracy and generalization rather than optimizing a single split.

2.7. Explainability and behavioral clustering

Model explainability was treated as a core component. SHAP quantified the contribution of each feature to model predictions [25], with global analysis identifying influential population-level predictors and local analysis interpreting individual predictions. To assess consistency, SHAP importances were recomputed across independent subsamples to test ranking stability. To capture latent behavioral patterns, a shallow autoencoder (input $\rightarrow$ 32 $\rightarrow$ 16 $\rightarrow$ 6 latent $\rightarrow$ 16 $\rightarrow$ 32 $\rightarrow$ output, trained with Adam and mean-squared-error loss for 30 epochs) compressed engineered behavioral features into a low-dimensional latent representation, which was then clustered using K-means. The number of clusters was determined by silhouette analysis over $k = 2$ to 10.

2.8. Evaluation metrics and validation

Model performance was assessed using mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). An 80/20 train/test split was applied, with the test set held out and never used during cross-validation, feature selection, or tuning. Five-fold cross-validation was performed on the training set only, and performance is reported as the mean with standard deviation across folds. All evaluation was internal to a single cohort; no separate external cohort was available, and the held-out split is therefore reported as internal validation rather than external validation. Further robustness analysis was conducted through sensitivity analysis (removing groups of related features) and an ablation study (removing individual features), both using the identical modeling pipeline so that results are directly comparable.

3. RESULTS AND DISCUSSION

3.1. Model performance comparison

The predictive performance of the framework was evaluated across five ML models using the reduced predictor configuration described in section 2.4. Table 3 summarizes the comparative results. Linear regression achieved a comparatively lower R^2 of 0.889, indicating that the relationship between the retained behavioral variables and the SHI is not fully captured by linear assumptions. Tree-based ensemble methods improved performance by modeling non-linear interactions: random forest reached $R^2 = 0.944$, while the gradient-boosting models LightGBM and XGBoost further improved consistency ($R^2 = 0.950$). Among all models, CatBoost provided the most stable and accurate performance ($R^2 = 0.951 \pm 0.001$); the low fold-to-fold variation indicates stable performance across the cross-validation folds. On the held-out test set, the final tuned CatBoost model achieved $R^2 = 0.951$, RMSE $= 0.013$, and MAE $= 0.011$.

Model robustness was further examined using five-fold cross-validation. As shown in Figure 2, fold-wise RMSE values remain consistently low with minimal variation across folds for all boosting models, indicating stable behavior across resampling rather than fold-specific artifacts. Overall, these results indicate that ensemble boosting models—particularly CatBoost—are well suited for modeling the continuous,

behavior-derived SHI. The performance is substantially below the near-perfect fit observed when SHI component variables are directly included, while remaining high under the reduced predictor configuration; the interpretation of this difference is examined in section 3.2.

Table 3. Comparative performance of ML models (five-fold cross-validation; mean ± standard deviation across folds)

Model	Category	R^2 (mean ± SD)	RMSE	MAE
Linear regression	Linear baseline	0.8892 ± 0.0017	0.02023	0.01581
Random forest	Bagging ensemble	0.9441 ± 0.0010	0.01437	0.01136
XGBoost	Gradient boosting	0.9499 ± 0.0007	0.01360	0.01082
LightGBM	Gradient boosting	0.9502 ± 0.0007	0.01356	0.01085
CatBoost	Ordered boosting	$\mathbf{0.9515 \pm 0.0007}$	**0.01338**	**0.01076**

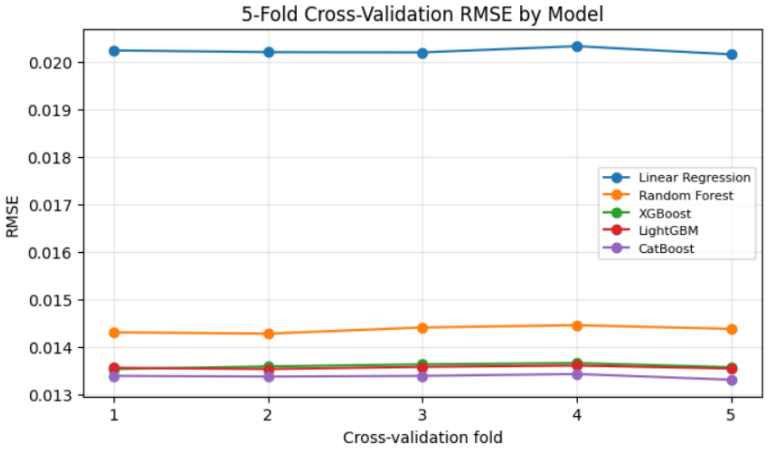

Figure 2. Five-fold cross-validation RMSE by model

3.2. Leakage demonstration and predictor-configuration analysis

To examine the effect of predictor configuration on measured performance, Table 4 contrasts two configurations on an identical data split. When the SHI components are included among the predictors, even a simple linear regression attains $R^2 = 1.000$; this is expected, since the index is by construction a linear combination of those components, and it is the signature of direct target leakage. When the SHI components are excluded (the proposed configuration), the linear baseline drops to $R^2 = 0.892$ while CatBoost reaches 0.952.

The leakage demonstration provides evidence that the proposed predictor configuration performs substantially below the deterministic reconstruction scenario in which SHI components are directly included. The comparison therefore demonstrates that the reduced predictor configuration substantially avoids the deterministic reconstruction observed when the SHI components are directly included. Because several retained behavioral variables remain related to SHI components—as detailed in section 2.4—the result is interpreted as predictive coherence rather than independent clinical outcome prediction, and not as formal construct validity. The gap between the linear and boosting results indicates that, once the components are removed, the remaining mapping is non-linear and is not recoverable by a linear reconstruction of the index; it does not establish that the model predicts an independently measured sleep-health outcome, and this table should not be read as demonstrating the complete absence of target-related information in the predictor set.

Table 4. Leakage demonstration on an identical data split

Predictor set	CatBoost R^2	Linear R^2
SHI components included (leaky)	0.99994	1.00000
SHI components excluded (proposed)	0.95203	0.89202

3.3. Feature selection results

The multi-stage feature-selection results are summarized in Table 5. Mutual information identified sleep debt, mid-sleep timing, wake time, and social jetlag as the most informative predictors, while Boruta confirmed a subset of six features as significantly relevant against randomized counterparts. The final retained subset consists of the Boruta-confirmed features that were also highly ranked by mutual information and SHAP: sleep debt, social jetlag, sleep-delay intensity, coping score, side-effect severity, and weekend sleep duration. Engineered behavioral scores were thus more informative for the model than raw contextual variables such as BMI or medical-condition status.

Table 5. Feature selection summary (Mutual information ranking and Boruta confirmation)

Feature	Mutual information	Boruta confirmed (retained)
sleep_debt	0.7361	Yes
midsleep_week	0.5113	No
wake_min	0.3589	No
social_jetlag	0.3121	Yes
bedtime_min	0.2215	No
sleep_delay_intensity	0.2016	Yes
coping_score	0.1943	Yes
side_effects_severity	0.1724	Yes
sleep_duration_weekend_est	0.1625	Yes
midsleep_weekend	0.1104	No
weekend_bedtime_min	0.0762	No
weekend_wake_min	0.0638	No
difficulty_falling_asleep	0.0148	No
concentration_difficulty	0.0039	No
restless_movements	0.0037	No
env_comfort_norm	0.0018	No
breathing_difficulty	0.0000	No
BMI	0.0000	No
has_medical_condition	0.0000	No

3.4. Global feature importance

To interpret the global decision behavior of the model, a SHAP-based feature-importance analysis was conducted. Figure 3 presents the SHAP summary (beeswarm) plot, illustrating both the relative importance of each behavioral feature and the direction of its association with model output. Sleep debt was the most influential predictor: higher sleep debt (red points) was associated with lower predicted sleep health. Wake time, sleep-delay intensity, and social jetlag followed as strong contributors, indicating that later and more variable scheduling was associated with lower predicted SHI. Side-effect severity and weekend sleep duration exerted moderate influence, whereas contextual variables such as BMI, breathing difficulty, and medical-condition status contributed little at the global level. These values describe how the model uses each feature and are associational; they do not establish causal effects on sleep health. This importance structure supports the design choice of emphasizing engineered behavioral features over raw questionnaire responses.

Because a common concern with post-hoc explanation is instability across resampling, the global SHAP ranking was recomputed on three independent subsamples of the test set. The top five features—sleep debt, wake time, sleep-delay intensity, social jetlag, and side-effect severity—were identical across all three subsamples. This repeated subsample analysis provides evidence of consistency in the ranking of influential predictors within the present cohort; it is not a complete validation of explanation stability.

3.5. Personalized SHAP analysis

To extend the global results to the individual level, a personalized SHAP analysis examined how specific behaviors jointly relate to the predicted SHI for a single subject. Figure 4 presents a local explanation for one individual, randomly selected from the test set and shown solely as an illustrative example rather than as a representative or typical case. The predicted score is decomposed into additive positive and negative contributions around the model's baseline. For this individual, elevated sleep debt and late-night behaviors contributed negatively to the prediction, while more regular timing contributed positively. The direction and magnitude of these contributions vary across individuals even when global rankings remain similar, so population-level importance does not translate uniformly to individual cases and no population-wide inference should be drawn

from this single example. This personalized decomposition enables fine-grained identification of the behaviors driving each individual prediction, providing a basis for targeted, exploratory decision-support guidance.

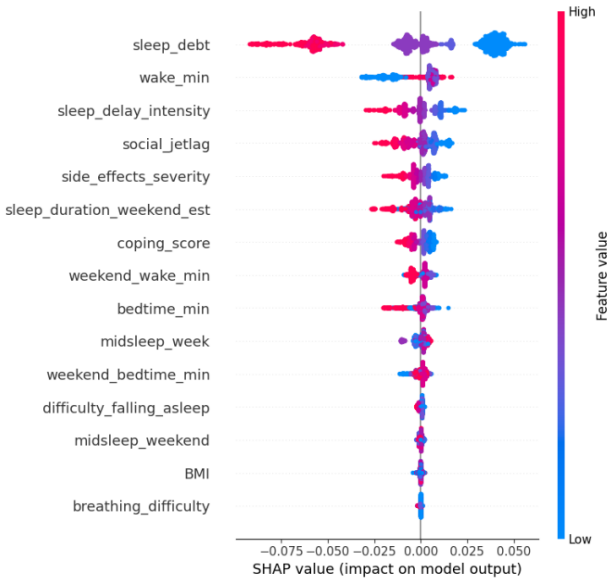

Figure 3. SHAP summary plot showing the global contribution of engineered behavioral features to the SHI prediction

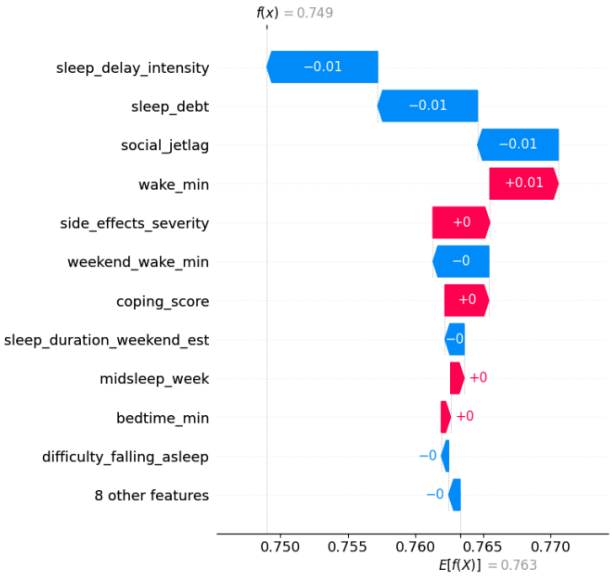

Figure 4. Personalized SHAP force plot illustrating individual-level contributions of behavioral features to the predicted SHI

3.6. Behavioral clustering results

To capture latent heterogeneity in sleep-related behaviors, an unsupervised clustering analysis was conducted using autoencoder-based representation followed by K-means. Silhouette analysis over $k = 2$ to 10 indicated that $k = 2$ yielded the strongest cluster separation, and a two-cluster solution was therefore selected. Table 6 reports the resulting cluster profiles. The first cluster (n=5146) is characterized by more stable routines, higher regularity and routine-consistency scores, lower social jetlag, and lower sleep debt. The second cluster (n=2854) exhibits reduced regularity and routine consistency, higher social jetlag, and substantially greater

sleep debt, indicating a more misaligned and sleep-deprived behavioral pattern. The resulting clusters exhibited distinct and interpretable behavioral profiles; however, these profiles should be regarded as data-driven groupings derived from the present cohort rather than externally validated behavioral archetypes. Combined with personalized SHAP explanations, they may support both individual-level guidance and group-level interpretation.

Table 6. Behavioral cluster profiles (mean values; $k = 2$)

Cluster	Bedtime (min)	Regularity index	Routine consistency	LNHI	Social jetlag	Sleep debt	Size
0 (more regular profile)	1453.47	0.92	0.91	0.41	90.25	66.30	5146
1 (more misaligned profile)	1450.71	0.87	0.85	0.31	141.56	105.91	2854

3.7. Sensitivity and ablation study

To assess robustness and internal consistency, sensitivity and ablation analyses were conducted using the identical modeling pipeline and the same held-out evaluation. In the sensitivity analysis, removing habit-related features (delay intensity, coping, and side effects) produced the largest performance degradation ($\Delta R^2 = -0.040$). Removing raw timing features produced a smaller drop ($\Delta R^2 = -0.007$), while removing social-jetlag, duration, symptom, or contextual groups produced negligible changes, suggesting these dimensions provide complementary rather than dominant information when jointly modeled. Feature-level ablation confirmed this pattern: sleep-delay intensity, side-effect severity, and coping score were the most critical individual contributors. Within the evaluated modeling framework, habit-related features therefore contributed more strongly to predictive performance than the other tested feature groups. This is a model-specific result, consistent with evidence that sleep-related behaviors are modifiable and behaviorally patterned [26], and should not be read as a general biological conclusion about the determinants of sleep health.

3.8. Comparison with existing studies

Table 7 positions the proposed framework against representative recent studies on machine-learning-based sleep assessment. Prior work predominantly targets categorical disorder detection or sleep-stage classification using physiological or wearable signals and reports limited interpretability; the present study instead models a continuous engineered index, reports an explicit leakage-control analysis, provides dual-level SHAP explanation with a consistency check, and adds behavioral clustering within a single unified framework. Because the target types differ—categorical stage or disorder labels in prior work versus a continuous engineered index here—the reported metrics are not directly comparable across rows, and the table is therefore organized by target type, data modality, explainability, personalization, and validation design rather than by headline performance values.

Table 7. Comparison with representative recent sleep-assessment studies. Metric column reports the type of metric used, not a comparable value; NR = not reported in a directly comparable form

Study	Data/target	Method	Explainability	Personalization	Validation design	Metric type
Stephansen *et al.* [4]	EEG/sleep stages	Neural network	No	No	Multi-cohort evaluation	Classification accuracy
Radha *et al.* [8]	HRV/sleep stages	LSTM	No	No	4-fold participant-level cross-validation	Classification accuracy, κ
El-Kenawy *et al.* [7]	Lifestyle/disorder	DTO + logistic reg.	Limited	No	Internal split	Classification accuracy
Khademi et al. [22]	Actigraphy/sleep parameters	Ensemble ML	No	Partial	NR	Error-based (NR)
Wang *et al.* [18]	Survey/disorder risk	LightGBM	SHAP (global)	No	Internal validation	AUC
Sathyanarayana *et al.* [15]	Wearable/sleep quality	Deep learning	No	No	NR	Classification accuracy
This study	Survey/continuous engineered SHI	CatBoost + AE-KMeans	SHAP (global + local, consistency check)	Yes (rule-based, exploratory)	Internal held-out split, single cohort	R^2, RMSE, MAE

3.9. Interpretation of the framework's four components

The framework comprises four components that differ in how strongly they are supported by the present evidence, and they should not be read as equally validated. Prediction is quantitatively evaluated: the model predicts the engineered SHI under the reduced predictor configuration with the performance reported in Table 3. Explanation is quantitatively and qualitatively examined: SHAP identifies the features contributing to model predictions, with a subsample consistency check, but remains associational and model-specific. Clustering is internally evaluated: the autoencoder–KMeans procedure yields data-driven behavioral profiles selected by silhouette analysis within this cohort. Recommendation is simulated and exploratory: the rule-based layer produces hypothetical decision-support guidance that has not been tested in any intervention.

3.10. Behavioral implications and computational feasibility

The framework translates explainable model outputs into structured sleep-health reasoning. Figure 5 presents a simulated illustration showing how the model responds to hypothetical adjustments in key behavioral variables and the corresponding inferred change in the predicted SHI (from 0.623 to 0.837 for the illustrated case). This visualization supports interpretability and decision-support reasoning only; it does not represent outcomes from a real-world intervention or longitudinal study, and no intervention efficacy is claimed or implied. From a behavioral perspective, the framework enables identification of the predictors most strongly associated with the modeled outcome—sleep debt, late-night habits, and social jetlag—without asserting that modifying them produces the corresponding change in measured sleep health. The framework is computationally feasible for exploratory analysis on standard computing resources, where all models are lightweight gradient-boosting or shallow architectures that train within minutes; detailed runtime, memory, scalability, and inference-latency profiling remains future work and was outside the scope of the present analysis. The unified integration of prediction, explanation, and rule-based guidance positions the framework as an exploratory decision-support tool, with empirical validation through longitudinal or interventional studies remaining necessary.

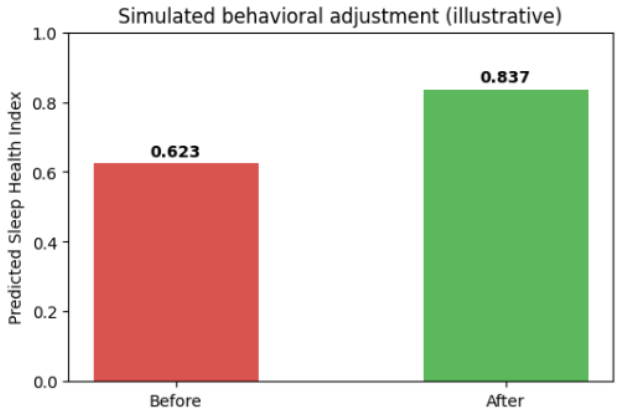

Figure 5. Simulated illustration of model-driven behavioral adjustment demonstrating the explanatory logic of the proposed framework. The before–after comparison reflects hypothetical changes in selected behavioral features and the corresponding SHI response inferred by the model. No real-world intervention, behavioral modification, or longitudinal validation is implied

3.11. Limitations and future work

Several limitations should be acknowledged. First, the dataset is based on self-reported survey responses, which are subject to recall bias and subjective interpretation; objective physiological signals such as EEG, actigraphy, or HRV were not incorporated. Recruitment was voluntary and convenience-based, the online and in-person proportions were not recorded, and the demographic representativeness of the sample could not be established, so self-selection bias cannot be excluded. Second, the SHI is an engineered composite construct rather than a clinically standardized diagnostic score and should be interpreted as a decision-support metric rather than a diagnostic substitute. Because the SHI is constructed from survey-derived behavioral components, high predictive performance should not be interpreted as evidence that the framework predicts

an independently measured clinical sleep-health outcome; moreover, as set out in section 2.4, several retained predictors remain structurally related to SHI components, so the configuration reduces but does not eliminate predictor–target overlap. Third, the evaluation was conducted on a single cohort without longitudinal follow-up; the held-out split constitutes internal validation only, the before–after illustration is simulated and explanatory rather than validated through intervention trials, and no external-cohort validation was performed. Detailed runtime, memory, and scalability profiling for large-scale deployment was also beyond the present scope. All source code and the processed dataset are publicly available to support transparency and reproducibility [27]. Future work will prioritize external and construct validation across diverse populations, integration of multimodal physiological data, longitudinal and interventional evaluation, and only subsequently the development of adaptive recommendation strategies.

4. CONCLUSION

This study proposed an explainable and personalized sleep-health assessment framework that integrates behavioral feature engineering, a composite SHI, and interpretable ML. By combining Mutual Information, Boruta, and SHAP-driven feature selection with an explicit leakage-control design, the framework makes the relationship between the engineered target and the predictor set transparent. The optimized CatBoost model achieved $R^2 = 0.951$ on a held-out test set under the reduced predictor configuration. The leakage demonstration showed a substantial reduction in predictive performance when the SHI component variables themselves were excluded, indicating that the proposed configuration does not reproduce the near-perfect performance observed under direct component inclusion. Nevertheless, because some retained behavioral variables are related to the SHI construction, the reported performance should be interpreted as prediction of an engineered behavioral index rather than independent validation of a clinical sleep-health construct.

Beyond predictive performance, the contribution lies in explainability and personalization: global and local SHAP analyses enabled transparent identification of the behaviors associated with higher and lower predicted sleep health—primarily sleep debt, late-night habits, and social jetlag—while behavioral clustering revealed interpretable, data-driven behavioral profiles. The strongest contribution of this work is the integrated analytical framework rather than a new learning algorithm or a clinically validated index. The framework provides an interpretable and exploratory approach for behavioral sleep-health assessment, while external, longitudinal, objective physiological, and real-world intervention validation remain necessary before clinical or deployment-level conclusions can be drawn.

ACKNOWLEDGMENTS

The author thanks all survey participants for their voluntary contribution to this study.

FUNDING INFORMATION

This research received no external funding; all costs related to data collection, processing, and analysis were personally funded by the author.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Neloy Pramanik Supto	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

The author declares no conflict of interest related to this study.

ETHICAL APPROVAL

Data collection used a structured questionnaire administered through Google Forms and face-to-face interactions in various districts of Bangladesh. Participation was entirely voluntary, and all respondents were informed in advance about the academic purpose of the research. Informed consent was obtained from every participant. Responses were gathered anonymously, with no personally identifiable or sensitive personal data recorded. In accordance with institutional and international research-ethics norms for non-clinical, observational studies involving anonymous survey data, no formal institutional review-board clearance was required, as the research did not involve medical treatment, clinical testing, or vulnerable groups. Only AI-assisted tools (ChatGPT and Grammarly) were used solely for language and grammatical improvement, and not during study design, data collection, analysis, modeling, interpretation, or scientific decision-making.

DATA AVAILABILITY

The data and source code that support the findings of this study are openly available in the GitHub repository at https://github.com/Neloy4321/Explainable-Sleep-Health [27], and the survey instrument is available at https://forms.gle/9vr4bMBKGtKvtzqt5 [23].

REFERENCES

[1] J. P. Chaput *et al.*, "Sleep timing, sleep consistency, and health in adults: a systematic review," *Applied physiology, nutrition, and metabolism = Physiologie appliquee, nutrition et metabolisme*, vol. 45, no. 10, pp. S232–S247, Oct. 2020, doi: 10.1139/apnm-2020-0032.

[2] S. L. Chellappa and D. Aeschbach, "Sleep and anxiety: from mechanisms to interventions," *Sleep Medicine Reviews*, vol. 61, p. 101583, Feb. 2022, doi: 10.1016/j.smrv.2021.101583.

[3] I. Perez-Pozuelo *et al.*, "The future of sleep health: a data-driven revolution in sleep science and medicine," *npj Digital Medicine*, vol. 3, no. 1, p. 42, Mar. 2020, doi: 10.1038/s41746-020-0244-4.

[4] J. B. Stephansen *et al.*, "Neural network analysis of sleep stages enables efficient diagnosis of NARCOLEPSY," *Nature Communications*, vol. 9, no. 1, p. 5229, Dec. 2018, doi: 10.1038/s41467-018-07229-3.

[5] S. Tiwari, D. Arora, and V. Nagar, "Supervised approach based sleep disorder detection using non - Linear dynamic features (NLDF) of EEG," *Measurement: Sensors*, vol. 24, p. 100469, Dec. 2022, doi: 10.1016/j.measen.2022.100469.

[6] V. Farrahi and M. Rostami, "Machine learning in physical activity, sedentary, and sleep behavior research," *Journal of Activity, Sedentary and Sleep Behaviors*, vol. 3, no. 1, p. 5, Jan. 2024, doi: 10.1186/s44167-024-00045-9.

[7] E. S. M. El-Kenawy, A. Ibrahim, A. A. Abdelhamid, N. Khodadadi, L. Abualigah, and M. M. Eid, "Predicting sleep disorders: leveraging sleep health and lifestyle data with dipper throated optimization algorithm for feature selection and logistic regression for classification," *Computational Journal of Mathematical and Statistical Sciences*, vol. 3, no. 2, pp. 341–358, Nov. 2024, doi: 10.21608/cjmss.2024.290167.1053.

[8] M. Radha *et al.*, "Sleep stage classification from heart-rate variability using long short-term memory neural networks," *Scientific Reports*, vol. 9, no. 1, p. 14149, Oct. 2019, doi: 10.1038/s41598-019-49703-y.

[9] N. Surantha, T. F. Lesmana, and S. M. Isa, "Sleep stage classification using extreme learning machine and particle swarm optimization for healthcare big data," *Journal of Big Data*, vol. 8, no. 1, p. 14, Dec. 2021, doi: 10.1186/s40537-020-00406-6.

[10] E. Dritsas and M. Trigka, "Utilizing multi-class classification methods for automated sleep disorder prediction," *Information (Switzerland)*, vol. 15, no. 8, p. 426, Jul. 2024, doi: 10.3390/info15080426.

[11] A. Smith, H. Anand, S. Milosavljevic, K. M. Rentschler, A. Pocivavsek, and H. Valafar, "Application of machine learning to sleep stage classification," in *Proceedings - 2021 International Conference on Computational Science and Computational Intelligence, CSCI 2021*, IEEE, Dec. 2021, pp. 349–354. doi: 10.1109/CSCI54926.2021.00130.

[12] C. Mencar *et al.*, "Application of machine learning to predict obstructive sleep apnea syndrome severity," *Health Informatics Journal*, vol. 26, no. 1, pp. 298–317, Mar. 2020, doi: 10.1177/1460458218824725.

[13] J. Palotti *et al.*, "Benchmark on a large cohort for sleep-wake classification with machine learning techniques," *npj Digital Medicine*, vol. 2, no. 1, p. 50, Jun. 2019, doi: 10.1038/s41746-019-0126-9.

[14] T. Wielek *et al.*, "Sleep in patients with disorders of consciousness characterized by means of machine learning," *PLoS ONE*, vol. 13, no. 1, p. e0190458, Jan. 2018, doi: 10.1371/journal.pone.0190458.

[15] A. Sathyanarayana *et al.*, "Sleep quality prediction from wearable data using deep learning," *JMIR mHealth and uHealth*, vol. 4, no. 4, p. e125, Nov. 2016, doi: 10.2196/mhealth.6562.

[16] E. Baghaeifar, S. Shamekhi, and F. Shalchizadeh, "An fNIRS-BCI study: effective channels selection in imagining right and left hand movements via brain functional connectivity," *Biomedical Signal Processing and Control*, vol. 109, p. 107915, Nov. 2025, doi: 10.1016/j.bspc.2025.107915.

[17] G. Atteia *et al.*, "Adaptive dynamic dipper throated optimization for feature selection in medical data," *Computers, Materials and Continua*, vol. 75, no. 1, pp. 1883–1900, 2023, doi: 10.32604/cmc.2023.031723.

[18] X. Wang *et al.*, "Development and validation of an explainable machine learning model for predicting the risk of sleep disorders in older adults with multimorbidity: a cross-sectional study," *Frontiers in Public Health*, vol. 13, Aug. 2025, doi: 10.3389/fpubh.2025.1619406.
[19] Y. Tang *et al.*, "Development of an explainable machine learning model for predicting depression in adults with type 2 diabetes mellitus: a cross-sectional SHAP-based analysis of NHANES 2009–2023," *Medicine (United States)*, vol. 105, no. 6, p. e47522, Feb. 2026, doi: 10.1097/MD.0000000000047522.
[20] D. Fischer *et al.*, "Irregular sleep and event schedules are associated with poorer self-reported well-being in US college students," *Sleep*, vol. 43, no. 6, pp. 1–12, Jun. 2020, doi: 10.1093/sleep/zsz300.
[21] O. Y. Azuara-Hernandez and Z. Gillette, "Using machine learning to determine optimal sleeping schedules of individual college students," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13521 LNCS, 2022, pp. 13–25. doi: 10.1007/978-3-031-17902-0_2.
[22] A. Khademi, Y. El-Manzalawy, L. Master, O. M. Buxton, and V. G. Honavar, "Personalized sleep parameters estimation from actigraphy: a machine learning approach," *Nature and Science of Sleep*, vol. 11, pp. 387–399, Dec. 2019, doi: 10.2147/NSS.S220716.
[23] N. P. Supto, "Sleep health behavior survey questionnaire," Google Forms, 2024. [Online]. Available: https://forms.gle/%0A9vr4bMBKGtKvtzqt5.
[24] K. Okano, J. R. Kaczmarzyk, N. Dave, J. D. E. Gabrieli, and J. C. Grossman, "Sleep quality, duration, and consistency are associated with better academic performance in college students," *npj Science of Learning*, vol. 4, no. 1, p. 16, Oct. 2019, doi: 10.1038/s41539-019-0055-z.
[25] S. Heddam, "Explainability of machine learning using shapley additive exPlanations (SHAP): CatBoost, XGBoost and LightGBM for total dissolved gas prediction," in *Studies in Big Data*, vol. 155, 2024, pp. 1–25. doi: 10.1007/978-3-031-66842-5_1.
[26] S. P. H. Khor, A. McClure, G. Aldridge, B. Bei, and M. B. H. Yap, "Modifiable parental factors in adolescent sleep: a systematic review and meta-analysis," *Sleep Medicine Reviews*, vol. 56, p. 101408, Apr. 2021, doi: 10.1016/j.smrv.2020.101408.
[27] N. P. Supto, "Explainable Sleep Health: Code and dataset repository," GitHub, 2024. [Online]. Available: https://github.com/Neloy4321/Explainable-Sleep-Healt.

BIOGRAPHIES OF AUTHORS



Neloy Pramanik Supto is an undergraduate student in the Department of Computer Science and Engineering at Daffodil International University, Dhaka, Bangladesh. His research interests include machine learning, explainable artificial intelligence, health informatics, and behavioral data modeling, with a particular focus on interpretable and personalized decision-support systems. His current work centers on developing explainable machine learning frameworks for behavioral sleep-health assessment. He can be contacted at email: pramanik15-4349@diu.edu.bd.