Prediction of Sleep Quality among Iraqi Female Students: Application of Interpretable Machine Learning Algorithms

Karam Abdullah [1*], Yahya Qasim Ibrahim Al-Fadhili [2]
[1-2] Department of Computer Science, College of Education for Pure Science, University of Mosul, 41002, Mosul, Nineveh, Iraq

* Corresponding Author: **Karam Abdullah**

Article Info

P-ISSN: 3051-3383
E-ISSN: 3051-3391
Volume: 07
Issue: 02
Received: 06-05-2026
Accepted: 04-06-2026
Published: 02-07-2026
Page No: 01-13

Abstract

Background: Sleep quality plays a crucial role in adolescents' health and well-being. The aims of this study is to predict the sleep quality of female students in Iraq by using interpretable machine learning algorithms.
Methods: A cross sectional data-driven study was carried out on 890 female students in Iraq in 2026. After data cleaning, variable preprocessing, and class balancing using the Synthetic Minority Oversampling Technique, three machine learning models were trained and tested for predicting sleep quality: Logistic Regression, Random Forest, and XGBoost. Model performance was evaluated by determining Accuracy, F1 score and Area Under the Receiver Operating Characteristic Curve (ROC-AUC). SHAP was used to explain and assess the contribution of the most influential factors associated with sleep quality, to boost the interpretability of the best-performing model.
Results: The machine learning models showed good performance in sleep quality prediction. In terms of the classification performance, among the evaluated models, XGBoost was the highest with accuracy 83.1% and macro F1-score 81.6%, and Random Forest was the highest with ROC–AUC 0.904 in terms of discriminative ability. SHAP analysis showed sleep duration, age, physical activity status, psychological domain of quality of life and mobile phone use as the most powerful predictors of sleep quality.
Conclusion: the results of this study confirm that the use of machine learning models in conjunction with interpretable techniques of artificial intelligence like SHAP has improved the predictive performance and also helped to establish the most important determinants of sleep quality. This strategy could help shape the creation of intelligent screening systems in the schools and tailored prevention interventions that will enhance the sleep health of adolescents.

DOI: https://doi.org/10.54660/IJAIET.2026.7.2.01-13

Keywords: Sleep Quality, Female, Students, Machine Learning

Introduction

Sleep is one of the most basic bio-needs of humans and a key component to physical, psychological and social health. Thus, it has a crucial role in brain development and growth, cognition, emotional regulation, integrity of the immune system, and improving overall quality of life [1, 2]. The National Sleep Foundation and American Academy of Sleep Medicine recommend that teenagers need 8 to 10 hours of good sleep per night. Most importantly, sleep quality is not just reduced to sleep length; it takes into account sleep onset latency, continuity, sleep efficiency, nocturnal awakenings, and daytime functioning. Therefore, the quality of sleep is considered as an important measure of overall health in adolescents [3]

In recent years, sleep disorders have become an important public health problem worldwide and sleep disturbance in adolescents is reported in many countries at a rate of 30-60% [4, 5]. Poor sleep is strongly linked to poor school performance, attention and cognitive learning problems, depression, anxiety, poor health, and reduced quality of life (QoL). Sleep quality is multi-faceted and depends on the person, their socioeconomic status, physical activity, the use of electronic devices, mental health, academic stress, and quality of life [6, 7]. These determinants have been well recognized, but the inter-relationships are very complex and are mainly non-linear, meaning that conventional statistical techniques often do not provide an accurate representation of these complex interactions.

Females are one of the most vulnerable subgroups with respect to sleep disorders. The various changes in hormones during the pubertal period, along with increased emotional sensitivity, academic stress, social pressures, burden of familial responsibilities, and higher rates of psychological distress among adolescent girls in comparison to boys, place adolescent girls at a heightened risk of poor sleep quality [8, 9]. This is supported by the literature in other countries, which invariably shows that insomnia rates and sleep duration are much higher in adolescent girls than in adolescent boys. This has made it a priority to address sleep disturbances and sleep health in adolescent girls both in the health care and education sector around the world [10, 11].

This is a very important issue in the context of Iraqi society. Iraq has experienced significant socio-political changes in the last two decades, where social insecurity, economic volatility, swift changes in lifestyle and lack of mental health infrastructure are constant features. These negative experiences have an adverse effect on the physical and psychological health of adolescents at the macro level and especially for adolescent girls. The multi-faceted financial burden on families, social fears, increased educational stress, limited recreational activities, growing use of digital devices and new lifestyle patterns all contribute to increased vulnerability of Iraqi female students to poor sleep hygiene [12, 12].

While many studies have examined the factors that influence good sleep quality in adolescents [14-16], most of these studies have used traditional statistical techniques, including linear and logistic regression models. These traditional methods have their drawbacks in understanding the complex, multidimensional and non-linear interactions between variables. In addition, empirical studies specifically focusing on Iraqi learners (especially females) are very limited. To the best of our knowledge, there has been no published extensive research that compares the predictive ability of contextual factors with a range of quality-of-life dimensions for this marginalized group in relation to their sleep quality.

The machine learning algorithms have been widely used in health sciences in recent years due to their ability of detecting complex patterns and enhancing the predictive accuracy in recent years. However, the models are inherently "black-box" and many of them are not easily interpretable. By using the interpretable machine learning (IML), highly accurate predictive models can be developed that also inform about how much each predictor contributes to the sleep quality estimate and how important each predictor is. This makes the results more interpretable, clinically meaningful and actionable for researchers, health workers and health policy makers [17, 18].

Conversely, quality of life, a broad measure of overall health, includes physical, psychological, social and academic functioning of adolescents and could be one of the most significant factors affecting sleep quality [19, 20]. However, it is not known to what extent the various components of quality of life, along with demographic and contextual factors, predict sleep quality among female students in Iraq, and how much more important are these components than the others in predicting high or low likelihood of a good night's sleep.

Hence, the main limitation in this research is the lack of a correct, understandable and machine learning-based predictive model that can detect the most important factors affecting the quality of sleep among Iraqi female students. In order to fill this gap, the current study utilizes interpretable machine learning algorithms to determine the most influential predictors of sleep quality, and to quantify the relative contributions of the demographic characteristics and the dimensions of the quality of life to the prediction of sleep quality.

Based on this, the first research question of this study is:

How well can demographic factors and the multiple dimensions of quality of life explain sleep quality among females in school in Iraq, using interpretable machine learning algorithms?

This research could help inform current knowledge about adolescent sleep health and serve as a scientific basis for designing new screening models, identifying youth who are at risk for poor sleep quality, developing early screening educational and psychological interventions, improving the quality of life, and their academic performance. Moreover, the results of this study can offer useful information to support decision making in the Iraqi Ministry of Health, the Ministry of Education, School administrators, educational counselor, educational psychologist, psychiatrist, sleep medicine doctor, public health researchers, health care planners, and educational policy makers for developing evidence-based prevention and intervention strategies.

Furthermore, the use of interpretable machine learning in this research could pave the way for a new research methodology in the future in the field of adolescent health and predictive medicine and may be useful in data-driven decision making in the healthcare and education systems in Iraq.

Methods

Study Design and Participants

The study was a cross-sectional study that was carried out in 2026 in the city of Mosul in Iraq. In the multistage sampling technique, 890 female students were enrolled in 10 schools of two education districts in Mosul.A total of 890 eligible female students were recruited using multistage sampling technique in 10 schools in two education districts in Mosul. The schools were randomly selected in each district and a simple random sampling technique was used within each school to sample student's representative of each grade level to complete questionnaires.

The students were selected according to the following criteria: female students, lower secondary school (grade 2 intermediate), and ability to understand the items of the questionnaire. Patients were excluded if they had a pre-existing medical condition/illness; and were not willing to participate in the study. Data collection began once permission was granted by the Ministry of Education and school authorities as well as written informed consent was obtained from school authorities and participants.

The data collection tool consisted of three parts. The first section collected demographic and background information, such as age, height, weight, parental level of education, parental occupation, family size, household income, sport clubs participation, physical activity engagement, hours of sleep per day, quality of sleep, outdoor physical activity and usage of smartphones each day. Sleep quality was the primary dependent variable and was divided into optimal vs. sub-optimal.

The third instrument was the 26-item World Health Organization Quality of Life (WHOQOL-BREF), which has been previously validated and reliable (34). This tool measures four areas: Physical health (7 items), Psychological health (6 items), Social relationships (3 items), and Environmental health (8 items). The domain scoring did not include item 1 (perception of overall quality of life) and item 2 (perception of health), as these items were examined separately. Answers were scored on a 5-point Likert scale (5 = very good, 1 = very bad). The total scores were classified as: 0–20 (poor), 21–40 (moderate), 41–80 (good/very good), and 81–100 (excellent) quality of life. The students were given questionnaires and reviewed after completion to check for blank data which were returned to the respondents for immediate completion. All students invited to participate were 100% responsive. The total average time of the survey was about 15-20 minutes.

In the first phase the information so collected, was carefully reviewed for completeness, quality of data entered, and consistency. Inconsistencies, missing data and implausible values were corrected or removed according to pre-established data cleaning procedures. After cleaning and quality checking, 890 valid samples were kept for final statistical analysis and model development with machine learning.

After preparing the data, different machine learning models were developed and tested using the final dataset. The robustness and reliability of the predictive outcomes were achieved by adopting standard data processing and standardization, partitioning training and testing sets, class balancing methods, and training multiple ML algorithms. A standard set of metrics were used to assess model performance. In addition, the most important predictors affecting sleep quality were selected using Explainable Artificial Intelligence (XAI) methods for improving the transparency and interpretability of the models.

Data Collection and Quality Control Procedure

At the end of the data collection period all questionnaires were carefully audited for completeness of answers, logical consistency and general data entry quality. The raw data was then imported into the analytical environment, where it underwent rigorous data quality control, data cleaning and data pipelines. After discarding incomplete and invalid data, a clean dataset of 890 valid observations was left for further statistical analysis downstream and to build predictive machine learning models.

Identify Variables and Measurement Instruments in The Study

Demographic variables were age, family size, household income, education of parents, and occupations of parents. Anthropometric measures included height, weight and Body Mass Index (BMI). BMI was determined using the formula:

$$\text{BMI} = \frac{\text{Weight (kg)}}{\text{Height (m)}^2}$$

Self-reported sleep quality and the length of the sleep period were analyzed as sleep-related variables. The sleep quality was used as the main target (dependent) variable in the machine learning models. In the binary classification task, it was dichotomized into two classes: optimal sleep quality (n=583) and sub-optimal sleep quality (n=307).

Other lifestyle and demographic factors were the duration of daily usage of the smartphone, social media engagement scores, physical exercise behavior, involvement in sports activities, membership in fitness clubs, and involvement in athletic associations. In addition, the participants quality of life was evaluated by the World Health Organization Quality of Life (WHOQOL-BREF) questionnaire, which consists of four domains: physical health, psychological health, social relationships and environmental health. Calculations of the scores for these domains were performed according to standard WHO scoring rules and used as input to the predictive models.

After data cleaning, quality control, categorical variable encoding and modeling preparation, 53 final features (numerical and categorical) were identified for building the machine learning models.

Data Preprocessing and Quality Control

Before the machine learning models were developed, a thorough data preprocessing pipeline was run that has been used to improve the quality of the data, to reduce measurement errors, and to prepare the dataset for computational analysis. All the pre-processing were performed in Python programming language (version 3.11) using specific data analysis and machine learning libraries.

The first step to the data analysis was to audit the raw data that was exported from the SPSS software. Quality Control methods involved missing value checks, duplicate checking, logical checking, and implausible value checks. Observations that contained data that were missing or were incorrectly recorded were corrected or omitted according to established standards. After this step, the final data set included 890 participants for further analyses down-stream.

Methods used to detect the outliers of numerical variables were range checking, Interquartile Range (IQR) method, and distribution analysis. Only in the presence of explicit evidence of data entry error was an outlier modified or corrected, and only when the omission of such data points from a model would significantly diminish its generalizability.

Due to the nature of the data (mixed numerical and categorical), feature transformation was conducted. One-Hot Encoding was used to transform categorical variables into numeric values appropriate for machine learning algorithms. In addition, numerical characteristics were standardized to reduce the scale differences which might affect the performance of the scale-sensitive algorithms, like Logistic Regression. The standardization process was carried out using the following formula:

$$Z = \frac{X-\mu}{\sigma}$$

Sleep-related variables evaluated were daily sleep duration and self-reported sleep quality. The sleep quality was used as the dependent (primary target) variable for the machine learning models. For the binary classification task, it was classified into two groups: optimal sleep quality (n = 583) and sub-optimal sleep quality (n = 307).

Other lifestyle and demographic factors consisted of average time spent using a smart phone per day, scores for social media use, physical exercise behaviors, sports activities participation, fitness club membership and athletic association membership. Additionally, participants' quality of life measured by the World Health Organization Quality of Life (WHOQOL-BREF) questionnaire on four core areas: physical health, psychological health, social relationships and environment. The scores for these domains were derived in line with the standard WHO scoring guidelines and used in the predictive models.

After data cleaning, quality control, categorical variable encoding, and modelling preparation, 53 final features (both numerical and categorical) were selected for developing the machine learning models.

The data was preprocessed and quality controlled.The data was preprocessed and quality controlled.

To ensure high-quality data, reduce measurement inaccuracies and prepare the data for computational analysis, a comprehensive data preprocessing pipeline was run before the development of the machine learning models.

The implementation of all the pre-processing steps has been done in Python programming language (version 3.11) with the help of some data analysis and machine learning libraries. The raw data from SPSS software were audited in the first phase. Quality control procedures involved checking for missing data, duplicate data, data with logical inconsistencies and extreme values. Incomplete data and/or recording errors were either corrected or removed by using predetermined criteria. After this, a down-streamed dataset of 890 participants was drawn up.

To detect any outliers in the numerical variables, statistical methods such as range verification, Interquartile Range (IQR) method, and distribution analysis were used. Data points were modified or corrected only when there was explicit evidence from the data points that they were data entry errors; otherwise, they were not corrected because the removal of valid data points can reduce the generalizability of predictive models.

The data was a mixture of numerical and categorical, so feature transformation was done. One-Hot Encoding technique was used to encode categorical variables into numerical values suitable for machine learning algorithms. Moreover, numerical features were standardized so that they are scaled between 0 and 1 to avoid scale differences causing bias in the performance of scale sensitive algorithms like Logistic Regression. Standardisation procedure was carried out as follows:

Note the notation: Z is the standardized value of the feature; X is the observed value of the variable; μ is the mean (average) of the variable; and σ is the standard deviation of the variable in the set.

To avoid leakage, all the parameters for feature transformation and feature standardization were calculated only on training data and then applied to the test data without any changes. This allows for the model to be evaluated in a way that is true to its predictive ability on new data. Thus, after the preparation process, a feature matrix, comprising 53 input features, was created and used to train the machine learning algorithms. The entire data preparation and model building process is summarized in Figure 1.

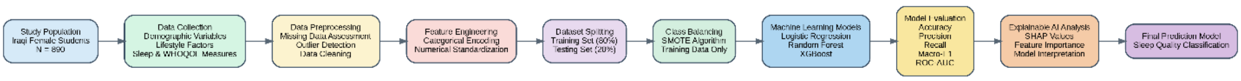

Fig 1: Machine Learning Workflow

Machine Learning Model Development and Interpretability Analysis

A supervised machine learning approach was employed to create predictive models of sleep quality, utilizing Python (version 3.11). Various specialized libraries such as scikit-learn, XGBoost, imbalanced-learn, and SHAP were used. This study was aimed at determining the best algorithm that Classifies Sleep Quality of Iraqi female students according to a wide range of demographic, anthropometric, physical activity, digital behavior, and quality of life parameters.

The final dataset of 890 participants was randomly divided into two sets by stratified split procedure, 80% being used to train the models and the remaining 20% for independent assessment of the performance of the models. Stratified partitioning was used to ensure that the distribution of sleep quality classes was similar for training and testing sets and reduce the possibility of class imbalance bias.

In the preprocessing step, the missing values and outliers were carefully checked and adjusted. To ensure that the numerical variables could be used by the machine learning algorithms, they were standardized. For categorical variables, they were encoded using One-Hot Encoding. Most importantly, the data did not leak into the model because all the preprocessing transformations were applied just to the training dataset; all testing data was processed after the training data was preprocessed.

The inherent class imbalance of the sleep quality target variable was accounted for by the synthetic minority oversampling technique (SMOTE), which is only applied to the training set. This method creates synthetic observations of the minority class to create a balanced class distribution when training. In contrast, the test data was kept intact, with no modifications, to ensure that the models' performance was evaluated in a unbiased and realistic way.

Three classifiers were developed and compared: Logistic Regression, Random Forest and Extreme Gradient Boosting (XGBoost). The choice of Logistic Regression as a baseline model was made because it is a statistically interpretable model, while Random Forest and XGBoost were chosen because they have good power to detect non-linearities and interactions between features. Each of the models was configured with different parameters and hyperparameters as detailed in Table 1.

The performance of the model was evaluated using a number of standard metrics: Accuracy, Sensitivity (Recall), Precision, F1-score, Macro-F1 average, Weighted-F1 average, and Area Under the Receiver Operating Characteristic curve (ROC-AUC).

To improve the transparency of the model and understand the reasons behind sleep quality, an Explainable Artificial Intelligence (XAI) technique, known as SHapley Additive exPlanations (SHAP), was used. The value of SHAP can be written as the sum of the feature importance contributions, allowing for a mathematically precise evaluation of the relative importance of each feature, which will help measure how vital it is for the model to make its final classification decision. In this way, feature importance was extracted, the decision-making process of the final model was made completely transparent, and the features most important to the sleep quality of students were ranked.

Table 1: Hyperparameters of machine learning models

Model	Hyperparameter	Setting
Logistic Regression	Maximum iterations	1000
	Solver	LBFGS
	Regularization	L2
	Random state	42
Random Forest	Number of trees	300
	Maximum depth	Default
	Criterion	Gini
	Random state	42
XGBoost	Number of estimators	300
	Learning rate	0.05
	Maximum depth	5
	Random state	42

Class Imbalance Solution: SMOTE

The Synthetic Minority Oversampling Technique (SMOTE) was used to equalize the distribution of the target variable (sleep quality) because of the class imbalance and ensure that the machine learning models do not become biased. In the case of classification models, an imbalance can have a huge impact on its ability to correctly classify minority class instances and thus lower the overall performance of classification algorithms. Therefore, it was necessary to use suitable oversampling methods to improve the learning of the model.

In this work, SMOTE algorithm was used exclusively on the training set and the independent test set was preserved as it is without modifications. This was specifically done to avoid leakage of data and also maintain the high validity of the final model evaluation. That is, no information in the unseen data was used in the generation of the synthetic data, and the testing data were only used to evaluate the performance of the trained models.

The dataset was first randomly split into a training dataset and a testing dataset, with 712 (80%) samples used for training and 178 (20%) samples used for testing. Later, the SMOTE algorithm was used to increase the number of minority class samples by creating new synthetic samples in the feature space so as to achieve a balanced class distribution. The training set was then increased to a balanced data set of 932 samples, with 466 and 466 samples from two sleep quality classes. The machine learning models such as Logistic regression, Random Forest and XGBoost were then trained using these balanced data.

Statistical Analysis and Model Evaluation

The statistical analysis and evaluation of the machine learning models were done to shed light on the characteristics of the participants' demographics, the performance of the algorithms, and the most accurate predictive algorithm for sleep quality in this study. The statistical analyses were performed using SPSS software (version 26.0) and the machine learning modeling and evaluation pipelines were carried out using Python (version 3.11).

Firstly descriptive statistics were used to describe the baseline data of the sample. For continuous variables, mean, standard deviation (SD), minimum, and maximum were provided. Frequencies and percentages were computed for categorical variables. The descriptive statistics of the participants are reported in Table 1.

In order to comprehensively evaluate the classification models, we calculated a number of standard evaluation metrics: Accuracy, Precision, Recall, F1 score, and Area Under the Receiver Operating Characteristic curve (ROC-AUC). These metrics were chosen as they possess good properties for assessing different facets of classification accuracy, especially when dealing with a binary classification problem with class imbalance.

Overall Model Accuracy: The percentage of instances that are classified correctly, which is computed as shown below:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

Precision quantifies the ratio of true positive predictions to the total positive predictions generated by the model, expressed as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

Recall (or Sensitivity) measures the model's capacity to correctly identify actual positive instances, formulated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

To establish a balance between Precision and Recall, the F1-score was computed as the harmonic mean of these two metrics:

$$\text{F1-score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}$$

Note: In the formulas above, TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives, respectively.

In addition, the ROC curve and the AUC values were computed to evaluate the discriminative power of the models in determining whether optimal or sub-optimal sleep quality groups. The AUC value is the likelihood that the model labels a randomly selected positive instance as positive, and a randomly selected negative instance as negative. A closer AUC to 1.0 means better discriminative capability while an AUC closer to 0.5 means performance similar to that of a random guesser.

In the final model selection, the predictive performance of the Logistic Regression, Random Forest, and XGBoost algorithms were extensively compared and contrasted, again using the same metrics. The best algorithm in terms of overall performance (Accuracy + F1 + AUC) was chosen as the best final model.

Explainable Artificial Intelligence (ExAI) with SHAP

An Explainable Artificial Intelligence (XAI) framework based on SHapley Additive exPlanations (SHAP) was applied to improve transparency, interpretability and comprehensibility of the machine learning decision-making pipeline. Despite having some of the best features in predicting non-linear and complex relationships among variables such as ensemble tree-based models like Random Forest and XGBoost, their internal decision-making processes appear as a "black box. In order to tackle this problem, SHAP decomposes the model predictions, quantifies the contribution of each input feature to the model prediction, and visualizes the results.

SHAP analysis was performed in this study for the best model that was selected using the highest predictive performance among the other models (Random Forest). For every observation and feature, SHAP values were calculated to show how much each variable contributed to the model's prediction. SHAP values were computed for each individual observation and feature, indicating the positive or negative effect each variable had on the final classification of the model. A SHAP value for a particular feature reflects the magnitude of the effect that the feature has on the model's prediction, and the further away from zero the larger the absolute value, the greater the relative importance of the feature in the model's decision-making process.

The theoretical basis for SHAP is based on the field of cooperative game theory (SHAPLEY values), which measures the contribution of each feature by considering all possible combinations of features (coalitions). The final model prediction for one case is represented by the additive feature attribution method:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i \, z'_i$$

where $g(z')$ represents the final model prediction in the given instance, ϕ_0 is the base value (the expected value or the mean prediction across the whole dataset), M is the total number of input features, and ϕ_i is the SHAP value for feature i, which indicates the contribution of the feature to the model's change in prediction for the given instance.

Statistical and Computational Environment

This particular study was completed using statistical and computational tools in the Python programming environment (version 3.11). All the process (data preparation, preprocessing, machine learning model development, evaluation, interpretability analysis of the model) was done through dedicated Python libraries. In particular, Pandas and NumPy were used for data manipulation and numerical processing, Scikit-learn for implementing the core machine learning algorithms and computing the evaluation metrics of performance, Imbalanced-learn for execution of SMOTE algorithm for class balancing, XGBoost for the creation of gradient-boosted tree models, and SHAP library for the Explainable Artificial Intelligence (XAI) and model interpretability analysis.

Ethical Considerations

The ethical principles of the Declaration of Helsinki and those of international ethical standards for human subject research were adhered to in this study. Before data were collected, formal approval and required clearances were gained with the relevant institutional scientific and ethical review boards. A clear explanation of the objective of the study, research methodology, nature of the data needed and the intended use of the data collected were provided to all participants and data collection commenced only after obtaining written informed consent.

All data collected was treated with strict confidentiality, and the data was anonymized extensively before going into the statistical analysis and machine learning modelling steps in order to protect the rights and privacy of the participants. Personally identifiable information was not used in the analysis, reporting of the results or in the writing of the manuscript. Moreover, the participants were explicitly told that they could withdraw from this study at any time, without any negative consequences. All subsequent data preprocessing, data cleaning, statistical computation and machine learning models development data pipelines were carried out on the anonymised version of the data only, to ensure absolute data privacy and uphold the highest standard of research ethics.

Results

Participant Characteristics and Descriptive Statistics

The mean age of participants was 16.03 ± 0.63 years, with an average BMI of 22.61 ± 3.85 kg/m². The mean sleep duration was 7.79 ± 1.81 hours per day, while daily mobile phone use averaged 3.95 ± 2.63 hours. Approximately one-third of participants reported gym participation (32.7%), whereas the majority were classified as physically inactive (87.9%). The mean scores of WHOQOL domains indicated moderate levels of quality of life among the participants. Table 1 summarizes the demographic, anthropometric, lifestyle, physical activity, and quality-of-life characteristics of the study participants.

Table 2: Demographic, anthropometric, lifestyle, physical activity, and quality-of-life characteristics of study participants (N = 890)

Variable	Category / Unit	Value
Demographic characteristics		
Age	Years, Mean ± SD	16.03 ± 0.63
Family size	Number of members,Mean ± SD	4.13 ± 0.69
Household income	Mean ± SD	19.27 ± 10.58
Anthropometric characteristics		
Height	m, Mean ± SD	1.65 ± 0.06
Weight	kg, Mean ± SD	61.75 ± 11.32
BMI	kg/m², Mean ± SD	22.61 ± 3.85
Sleep-related characteristics		
Sleep duration	hours/day, Mean ± SD	7.79 ± 1.81
Sleep quality classification		
Good sleep quality	n (%)	583 (65.5%)
Poor sleep quality	n (%)	307 (34.5%)
Digital behavior characteristics		
Mobile phone use	hours/day, Mean ± SD	3.95 ± 2.63
Social media use score	Mean ± SD	16.70 ± 5.09
Physical activity characteristics		
Gym participation		
Yes	n (%)	291 (32.7%)
No	n (%)	599 (67.3%)
Sport association membership		
Yes	n (%)	205 (23.0%)
No	n (%)	685 (77.0%)
Physical activity status		
Active	n (%)	108 (12.1%)
Inactive	n (%)	782 (87.9%)
Exercise status		
Regular exercise	n (%)	382 (42.9%)
Occasional exercise	n (%)	96 (10.8%)
Exercise preparation stage	n (%)	79 (8.9%)
No current exercise	n (%)	225 (25.3%)
Previous exercise	n (%)	108 (12.1%)
Quality of life characteristics (WHOQOL domains)		
WHOQOL Physical domain	Mean ± SD	6.39 ± 1.38
WHOQOL Psychological domain	Mean ± SD	3.70 ± 0.68
WHOQOL Social domain	Mean ± SD	3.63 ± 0.84
WHOQOL Environment domain	Mean ± SD	3.63 ± 0.75

Footnote

Values are presented as mean ± standard deviation (SD) or number (%). BMI: body mass index; WHOQOL: World Health Organization Quality of Life assessment.

Distribution of Sleep Quality Classes

In this study, the main study endpoint was sleep quality, divided into two clinical groups. The 890 participants evaluated were divided into two groups: optimal sleep quality (583, 65.5%) and sub-optimal sleep quality (307, 34.5%). The absolute necessity of using advanced data-balancing methods, such as the Synthetic Minority Oversampling Technique (SMOTE), during the computational phase, to avoid a predictive bias toward the dominant class and achieve good generalization performance was revealed.

Data Balancing using SMOTE

The distribution of the baseline was also strongly imbalanced, meaning that an untuned model would have a "training bias" towards the majority class, significantly limiting its ability to detect instances of the minority class. The Synthetic Minority Oversampling Technique (SMOTE) was applied to handle the issue of class imbalance and to improve the performance of the models trained on the downstream machine learning models in terms of generalization.

Importantly, the SMOTE algorithm was applied only in the training partition (80% of the original data set) to avoid data leakage and adding synthetic artifacts that contain characteristics of the test set. SMOTE works by generating new, artificial samples for the minority class instead of by just copying samples. These new synthetic observations are created on line segments connecting any and all of the k-nearest neighbors in the feature space.

The training set had a total of 712 observations before oversampling. After applying SMOTE the training set became a perfectly balanced subset of 932 observations (466 optimal, 466 sub-optimal). This even-weighted training process made the machine learning algorithms learn the predictive patterns in both groups equally, leading to the best classification performance. This optimised and balanced training matrix was then used to train the Logistic Regression, Random Forest and XGBoost algorithms which were then tested against standard metrics such as Accuracy, Precision, Recall, F1 score and ROC-AUC.

Table 3: Distribution of sleep quality classes before and after SMOTE balancing

Sleep quality class	Before SMOTE	After SMOTE
Class 1	466	466
Class 2	246	466
Total	712	932

Machine Learning Model Performance

Table 3 compares the predictive performance of the Logistic Regression, Random Forest and XGBoost algorithms in predicting sleep quality. The overall performance of the ensemble tree-based models was systematically better than the linear baseline model (Logistic Regression). Logistic Regression had the lowest prediction ability with an Accuracy of 69.1% and an ROC-AUC of 0.722.

The overall performance of the XGBoost model was the best in terms of the classification accuracy and balanced metrics, with Accuracy of 83.1% and Macro F1-score of 81.6%. The RF model had a slightly lower classification accuracy than the XGBoost model, but had the highest discriminative power with an ROC-AUC score of 0.904. This suggests that the Random Forest architecture has a better ability to accurately distinguish between the optimal sleep quality and sub-optimal sleep quality groups in the study sample.

Table 4: Performance comparison of machine learning models

Model	Accuracy	Macro F1-score	Weighted F1-score	ROC-AUC
Logistic Regression	0.691	0.663	0.694	0.722
Random Forest	0.826	0.804	0.825	0.904
XGBoost	0.831	0.816	0.833	0.883

The Performance of The Random Forest Model in Classification

Random Forest model turned out to be the best discriminative power with the highest ROC-AUC score and was chosen for detailed classification analysis. The performance of this model on the independent testing data set is reported in detail in Table 4.

The Random Forest model was able to get an overall Accuracy of 82.6% on the unseen test data. The model achieved the best results (good quality class) with a Precision of 0.86, a Recall of 0.88, and an F1-score of 0.87. The model achieved Precision of 0.76, Recall of 0.72, and F1-score of 0.74 for both classification targets for the sub-optimal (poor) sleep quality class, showing a good and balanced performance across both targets.

Table 5: Classification report of Random Forest model

Class	Precision	Recall	F1-score	Support
Better sleep quality	0.86	0.88	0.87	117
Poorer sleep quality	0.76	0.72	0.74	61
Accuracy			0.83	178
Macro average	0.81	0.80	0.80	178
Weighted average	0.82	0.83	0.82	178

ROC Curve Analysis

A Receiver Operating Characteristic (ROC) curve was carried out to test the ability of the machine learning models developed in this study for accurate discrimination between the two sleep quality groups. ROC curve depicts the relationship between TPR (Sensitivity) and FPR (1-Specificity) with respect to different levels of decision thresholds. In addition, Area Under the ROC Curve (AUC) was calculated as a threshold independent measure, which represents overall discriminative power of all models.

The outcomes of the ROC analysis showed that the developed models had strong predictive powers in the classification of students' sleep quality. The highest discriminative capacity (AUC = 0.904) was found in the Random Forest algorithm among the evaluated algorithms. This is indicative of a greater predictive value of identifying sleep quality profiles. The XGBoost model also performed well with an AUC score value of 0.883, while the base model of Logistic Regression performed the poorest in discriminative power ("AUC" = 0.722) among the ensemble tree-based architectures.

The results indicate that non-linear ensemble approaches, such as Random Forest (RF), are better suited for sleep quality classification than linear approaches, like Logistic Regression (LR), because of their natural capacity to model complex covariate space interactions and non-linear boundaries. Therefore, the Random Forest model was chosen as the final optimal model for further interpretability analysis and extracting feature importance from the SHAP framework. The comparative ROC curves and AUC values of the three machine learning models are shown in Figure 2.

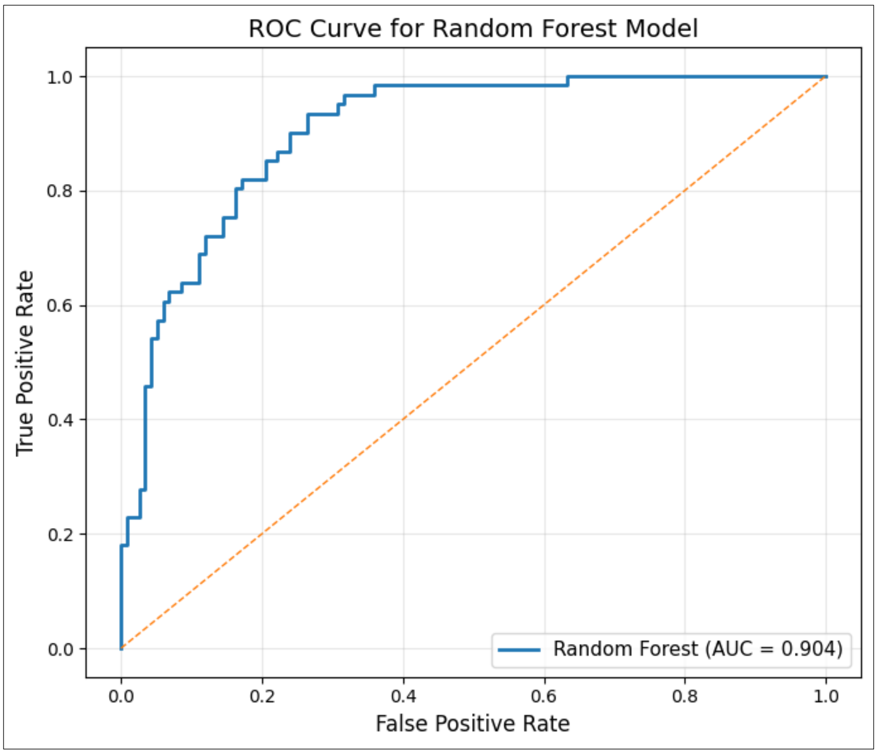

Fig 2: Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values for the evaluated machine learning models

Explainable Artificial Intelligence Analysis (SHAP)

To decode the decision-making process of the selected optimal model and identify the most critical determinants of sleep quality predictions, an Explainable Artificial Intelligence (XAI) analysis was performed on the final Random Forest model using the SHAP (SHapley Additive exPlanations) framework. Although machine learning architectures excel at identifying complex, non-linear interactions among covariates, their internal mechanics are inherently difficult to interpret directly. Consequently, deploying the SHAP approach enabled a rigorous evaluation of each feature's marginal contribution to the model's output, thereby enhancing the transparency and interpretability of the predictive pipeline.

The SHAP analysis revealed that daily sleep duration was the single most pivotal predictor of sleep quality, capturing the highest importance value among all input features. This was sequentially followed by age, exercise status, the psychological domain of quality of life (WHOQOL Psychological), and daily mobile phone usage, which collectively exerted the most substantial influence on the Random Forest model's classification decisions.

Furthermore, the results underscored that lifestyle factors and daily behavioral routines—specifically physical exercise engagement and digital technology usage—alongside psychological dimensions of quality of life, play an instrumental role in stratifying students' sleep quality. These insights indicate that predicting sleep quality extends beyond baseline sleep architecture and is multi-factorially driven by a complex interplay of individual, behavioral, and psychological covariates.

The global feature importance ranking, derived from the mean absolute SHAP values (|SHAP value|), demonstrated that features with higher values contributed more heavily to shifts in the model's output. In other words, these variables had the most profound impact on driving the model's probability estimates toward either optimal or sub-optimal sleep quality classifications. The comprehensive global feature importance plot based on SHAP values is illustrated in Figure 3.

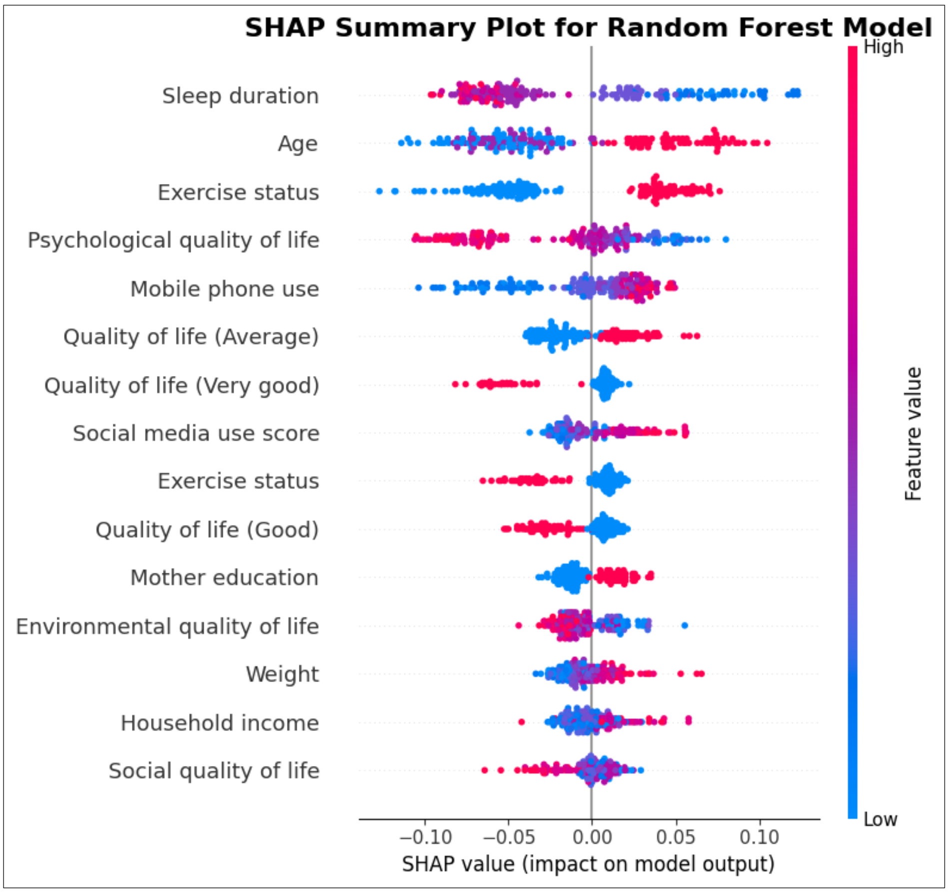

Fig 3: SHAP-based feature importance ranking for the Random Forest model in sleep quality prediction.

Discussion

The present study revealed machine learning algorithms and especially ensemble learning methods are very effective in sleep quality prediction for adolescent girls in Iraq. The high classification performance of XGBoost and the high discriminative power of Random Forest implies multi-dimensional and non-linear interactions between the demographic, behavioural and psychological features and suggest that sleep quality is determined by complex interactions. The results are in line with many studies showing that tree-based ensemble algorithms are more successful than traditional statistical models in predicting health outcomes involving multiple factors due to their ability to identify nonlinearity among predictors, higher-order interactions, and complex interactions between predictors [21, 22]. Logistic regression, in contrast, showed lower performance, suggesting that linear relationships between predictors and sleep quality may not be an appropriate model for understanding sleep health in adolescence, which may be influenced by a complex network of processes. [21, 23]. During adolescence, there are significant developmental shifts in circadian rhythm, greater academic requirements and increased social involvement, which all tend to lead to later sleep onset and shorter sleep duration [24]. Therefore, length of sleep is a basic indicator of sleep health. This variable was also identified as the most powerful predictor, further emphasizing the importance of explainable machine learning models for early identification of adolescents who are more likely to suffer from poor sleep quality [25].

Even though the age range of the study population was fairly limited, age also proved to be one of the most crucial predictors. This finding may be due to developmental changes during the adolescent period such as hormonal changes, delayed circadian phase preference, and increased school workloads and freedom in sleep timing. Previous research has shown that sleep patterns worsen over time during this period of development in adolescence, and that older adolescents have shorter sleep duration and poorer sleep quality. Thus, the age contribution in the current model could reflect the aggregate of these life course changes in biological and behavioral functioning [26, 27].

The other predictor variables identified by the model were physical activity and. This is consistent with prior epidemiological data indicating that physical activity has a variety of physiological and psychological mechanisms that lead to healthy sleep, such as regulation of the circadian rhythm, stress reduction, improved emotional state, and increased sleep efficiency. On the other hand, lack of exercise has been linked to lower sleep quality in adolescents on many occasions. The cross sectional design does not permit causal inference, but the contribution of physical activity to the prediction indicates that physical activity may be a modifiable behavioral risk factor that can be included in school-based sleep health promotion interventions [28, 29].

In addition, the psychological domain of quality of life was also found to be a significant factor for predicting sleep quality. This is corroboration of the known bidirectional relationship between psychology and sleep. Psychological distress, anxiety, depression and diminished life satisfaction can disrupt sleep initiation and maintenance, leading to greater cognitive and emotional arousal, while sleep disturbances can contribute to emotional dysregulation, further influencing psychological dysfunction. The results of these studies emphasize the need to incorporate mental health promotion into programs aimed at improving sleep quality among adolescents, not to treat sleep problems alone [30, 31].

In addition, the use of mobile phones and social media were found to be significant predictors of sleep quality. This is consistent with growing research suggesting that overuse of smartphone and social media, especially in the evening, can negatively impact sleep for multiple reasons, such as exposure to blue light, a delay in the onset of melatonin, longer time to fall asleep, cognitive over-stimulation, and displacement of sleep time. With adolescents' widespread use of digital technology, these findings highlight the need to foster healthy digital behaviors and to use educational strategies that decrease screen time at night as part of comprehensive sleep health strategies [32, 33].

The current study's main advantage is the seamless fusion of robust machine learning models and explainable AI (XAI) methods. Previous predictive studies have mainly aimed at optimizing the accuracy of models, but in this study, the use of SHAP significantly enhanced the interpretability of the models by providing a quantitative measure of the impact of each predictor on the model's predictions. This degree of transparency makes prediction models more clinically and public health applicable, allowing health professionals and school administrators to have a greater understanding of the factors that contribute to poor sleep. Therefore, explainable machine learning has a great potential to help develop intelligent screening systems in schools and personalized preventive interventions for improving adolescent sleep health.

Limitation

Although the methods used in this study are interpretable machine learning algorithms and a new predictive framework for sleep quality, there are some limitations that need to be taken into account when interpreting the results and generalizing them. First, the cross-sectional design does not allow drawing causal inferences with regard to the relationships between the demographic indicators, quality of life and sleep quality. The interpretations of the findings should thus be taken as associations rather than cause-and-effect. Second, all the study data were obtained by self-report questionnaires, which may be susceptible to recall bias, social desirability bias, and the different interpretation of the items of the questionnaires by the respondents, all of which could influence the accuracy of the data. Third, sleep quality was measured using self-reported measures not with objective methods like polysomnography (PSG) or actigraphy. Therefore, the results obtained are largely based on participants' self-reported sleep quality, rather than on the physiological aspects of sleep.

Fourth, the study was limited to female students in schools in Iraq. Thus, care needs to be taken when extrapolating the results to male students, out-of-school adolescents or other countries where cultural, social and socioeconomic differences could affect sleep patterns and the factors associated with them. Fifth, demographic factors and quality of life were the most important factors included in the final predictive model, but other factors that may influence sleep quality such as a genetic predisposition, clinician diagnosed sleep disorders, medication use, caffeine intake, dietary habits, objectively measured physical activity, exposure to blue light, sleep environment characteristics and family or social stressors were not included in the final model. Therefore there is a potential for residual confounding which cannot be ruled out.

Sixth, machine learning algorithms can represent complex and nonlinear relationships between predictors, but the quality, quantity and balance of training data can still substantially affect the performance of the machine learning algorithms. This imbalance in the data set, missing values or the small sample size can therefore limit predictive performance and model generalizability. Seventh, although interpretable machine learning can help provide transparency by revealing the relative importance of variables that are used to predict the model, feature importance and patterns generated from the model should not be interpreted as causalities. The results need to be confirmed by further studies, especially longitudinal and interventional.

Despite these shortcomings, this research has several significant merits. The combination of interpretable machine learning tools with a thorough analysis of demographic characteristics and multidimensional quality-of-life domains, combined with the study of an under-researched population—females in the context of Iraq—offers a powerful and novel approach to the study of sleep quality. The results can inform future research and evidence-based screening, interventions, and policy development to promote adolescent sleep health.

Conclusion

It improved sleep quality in adolescent girls while providing informative analysis of factors affecting sleep health via the interpretable artificial intelligence (AI) methods of SHAP. These findings underline the complexity of sleep quality and the need for considering behavioral and psychosocial factors in predictive models, with sleep duration, age, physical activity, psychological well-being and mobile phone use being identified as the most influential factors. The implications for practice are important with regard to the creation of data-driven, school-based screening systems and personalized preventive interventions that will allow the early identification of at-risk adolescents for poor sleep quality and for targeted health promotion interventions. However, the cross-sectional study design does not allow for causal inferences to be drawn, and the results need to be interpreted with that in mind. To further validate these models in larger, more diverse populations and include longitudinal and multicenter data to assess temporal changes and model generalizability, and to explore the feasibility of using additional objective data, such as data from wearable devices and digital behavioral measures, to increase the predictive accuracy and to implement EAI in adolescent health settings.

Acknowledgements

The authors would like to express their sincere gratitude to all the female students who voluntarily participated in this study. Their valuable time, cooperation, and willingness to share their experiences made this research possible. The authors also appreciate the support of the participating schools, teachers, and educational authorities who facilitated the data collection process. Their contributions are greatly appreciated.

Consent for publication

Not Applicable

Competing interests

The authors declare that they have no competing interests

Funding
The authors have not received any financial support for the research.

References

1. Benýšková E, Beníčková M, Gimunová M. The effect of sleep deprivation on cognitive function in postmenopausal women: a systematic review. Int J Womens Health. 2026;18:562840.
2. Jabade M, Nadaf H, Kaur J, Singh S, Deshmukh J. Unlocking the link: sleep deprivation and cognitive performance in working-age women: a systematic review. Unnes J Public Health. 2025;14(1):1-13.
3. Krempel R, Jarvers I, Ecker A, Schleicher D, Brunner R, Kandsperger S. Sleep quality and the cortisol and alpha-amylase awakening responses in adolescents with depressive disorders. BJPsych Open. 2024;10(5):e140.
4. Jani S, Parhiz A, Aminzadeh A, Javanmardi K, Jafarinezhad F. Worldwide prevalence of poor sleep quality in adolescents aged 10-19 years: a systematic review and meta-analysis. Iran J Child Neurol. 2026;20(2):1-11.
5. Li D, Nie X, Li J, Tao Y, Zhao C, Zhong H, *et al*. Factors associated with sleep disorders among adolescent students in rural areas of China. Front Public Health. 2023;11:1152151.
6. Zhou J, Liu Y, Yue C, Wang M, Chen K, Rosales K. The effect of sleep quality on academic performance: a systematic review and meta-analysis. Behav Sci. 2026;16(5):634.
7. Majeed A, Khudija Tul K, Asmara S, Gulrukh R, Shaikh Khalid M, Muhammad N. Exploring the relationship between sleep quality and cognitive performance among university students experiencing academic stress. Insights J Life Soc Sci. 2025;3(9):108-115.
8. Kaur H, Hasoglu T, Gomaa H, Ricci A, Criley C, Calhoun S, *et al*. Sleep disparities in adolescent women: role of pubertal development, menstrual cycle and premenstrual symptoms. Sleep. 2021;44(Suppl 2):A72-A73.
9. Meneo D, Baldi E, Benz F, Fabris M. A biopsychosocial approach to sleep health during puberty: individual and contextual aspects and the role of gender differences. A narrative review. Cogent Ment Health. 2025;4(1):2541697.
10. Zhang J, Chan NY, Lam SP, Li SX, Liu Y, Chan JWY, *et al*. Emergence of sex differences in insomnia symptoms in adolescents: a large-scale school-based study. Sleep. 2016;39(8):1563-1570.
11. de Zambotti M, Goldstone A, Colrain I, Baker F. Insomnia disorder in adolescence: diagnosis, impact, and treatment. Sleep Med Rev. 2018;39:12-24.
12. Tufeq ZA, Ziyab AM, Ahmed BA, Rahho MKN, Abdulah DM. Sleep patterns, insomnia, and well-being among female high school students: a population-based study. Sleep Med Res. 2023;14(2):105-112.
13. Al Shawi A, Tayib Z, Lafta R, *et al*. Insomnia, a sequel of stress among medical students in Iraq. F1000Research. 2026;15:258.
14. Maratia F, Bacaro V, Crocetti E. Sleep is a family affair: a systematic review and meta-analysis of longitudinal studies on the interplay between adolescents' sleep and family factors. Int J Environ Res Public Health. 2023;20:4572.
15. Bobba B, Bacaro V, Crocetti E. Embedded in contexts: a systematic review of the longitudinal associations between contextual factors and sleep. Adolesc Res Rev. 2023;8:403-422.
16. Liu D, He J, Li H. The relationship between adolescents' physical activity, circadian rhythms, and sleep. Front Psychiatry. 2024;15:1415985.
17. Anik MA, Topu MM, Wasi A, Khan MI, Ahsan MM. Integrating explainable machine learning and mixed-integer optimization for personalized sleep quality intervention. arXiv. 2026;2603.16937.
18. Kim M, Yun S, Kim H, Matsushita E, Yeom J, Kim S, *et al*. Heart rate circadian phase as wearable digital phenotyping of insomnia: an interpretable machine learning study. Digit Health. 2026;12:20552076261458929.
19. Hay CM. The relationship between sleep quality and quality of life in adolescents aged 16-18 years: a cross-sectional study. Sleep Med Res. 2025;16(4):241-247.
20. Ahmadi Z, Omidvar S. The quality of sleep and daytime sleepiness and their association with quality of school life and school achievement among students. J Educ Health Promot. 2025;11(1):159.
21. Zhang L, Zhao S, Yang W, Yang Z, Wu Z, Zheng H, *et al*. Utilizing machine learning techniques to identify severe sleep disturbances in Chinese adolescents: an analysis of lifestyle, physical activity, and psychological factors. Front Psychiatry. 2024;14:1447281.
22. Alghwiri A, Almomani F, Alghwiri A, Whitney S. Predictors of sleep quality among university students: the use of advanced machine learning techniques. Sleep Breath. 2021;25(2):1119-1126.
23. Wang W, Zhu Y, Yu H, Wu C, Li T, Ji C, *et al*. The impact of sleep quality on emotion regulation difficulties in adolescents: a chained mediation model involving daytime dysfunction, social exclusion, and self-control. BMC Public Health. 2024;24:1862.
24. Jiang M, He Z. A lifestyle-based prediction model for obesity in Chinese adolescent students. Front Sports Act Living. 2025;7:1677707.
25. Liangyu Z, Wenze S, Yan G, Zhang S, Yuke Y, Sen M, *et al*. Machine learning-based prediction of adolescents' quality of life. Int J Phys Act Health. 2026;5(1):284.
26. Singh R, Suri J, Sharma R, Suri T, Adhikari T. Sleep pattern of adolescents in a school in Delhi, India: impact on their mood and academic performance. Indian J Pediatr. 2018;85(10):841-848.
27. Kortesoja L, Vainikainen M, Hotulainen R, Rimpelä A, Dobewall H, Lindfors P, *et al*. Bidirectional relationship of sleep with emotional and behavioral difficulties: a five-year follow-up of Finnish adolescents. J Youth Adolesc. 2020;49(6):1277-1291.
28. Song H, Jiang X, Li Y, Cao Y, He C, Zuo H, *et al*. Effects of physical activity on sleep in children and adolescents: a systematic review and meta-analysis of randomised controlled trials. Eur Child Adolesc Psychiatry. 2026;35(2):397-411.
29. Yan H, Xin G, Chen R. Effects of exercise interventions on sleep quality in adolescents: a systematic review and meta-analysis. Front Public Health. 2025;13:1623506.
30. Zhai K, Gao X, Wang G. The role of sleep quality in the psychological well-being of final-year undergraduate students in China. Int J Environ Res Public Health. 2018;15(12):2881.

31. Klyne D, Hall M, Smith S, Barbe M, Egorova-Brumley N, Serafimovska A, *et al*. Understanding the relationship between sleep, psychological and musculoskeletal health from a neuroimmune perspective. Musculoskelet Sci Pract. 2026;82:103509.
32. Fatima G, Shukla V, Magomedova A, Magomedov R. Mobile phone addiction and sleep quality among children and adolescents: unraveling the health consequences. Cureus. 2026;18(1):e102460.
33. Demirci K, Akgönül M, Akpinar A. Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students. J Behav Addict. 2015;4(2):85-92.

How to Cite This Article
Abdullah K, Al-Fadhili YQI. Prediction of sleep quality among Iraqi female students: application of interpretable machine learning algorithms. Int J Artif Intell Eng Transform. 2026;7(2):1-13.
doi:10.54660/IJAIET.2026.7.2.01-13.