PREDICTING SLEEP QUALITY VIA BEHAVIOURAL AND LIFESTYLE INDICATOR ANALYSIS USING MACHINE LEARNING

Mr. Anand Mohan Shenoy,
Master of Science in Information Technology Part-I,
SIES (Nerul) College of Arts, Science & Commerce (Autonomous), Navi Mumbai, Maharashtra, India,
Email address: anandshenoy29@gmail.com

ABSTRACT

Sleep deprivation has become a critical public health concern, increasingly driven by modern lifestyle habits such as excessive screen time and high stress levels. While traditional clinical studies often focus on medical disorders like Sleep Apnoea, this research shifts the focus to predicting sleep quality by analysing behavioural and lifestyle indicators using machine learning. A primary dataset was collected via a structured survey capturing key metrics including sleep duration, latency, and bedtime procrastination, with a specific demographic emphasis on Gen Z and Millennials to reflect digital-age habits. A comparative analysis of six machine learning classifiers ranging from linear models to ensemble methods was performed and optimised using GridSearchCV. To ensure model reliability, the study employed a rigorous validation framework, introducing a custom "Overfitting Gap" threshold to guarantee generalisability. The results demonstrate that the Random Forest classifier achieved superior performance with 85.19% accuracy and an ROC-AUC of 0.88, proving more effective at handling complex, non-linear lifestyle data than traditional linear models. Crucially, Feature Importance analysis identified Sleep Duration, Bedtime Procrastination and Stress Level as the most significant predictors, highlighting that psychological behavioural patterns are as critical as biological markers in determining sleep health. This study provides a data-driven basis for personalised health interventions and future mobile-based sleep monitoring systems.

Keywords: Bedtime Procrastination, GridSearchCV, Machine Learning, Overfitting Gap, Random Forest, Sleep Deprivation

1. INTRODUCTION

In the digital age, sleep patterns have shifted drastically due to technological ubiquity and changing social norms. While often associated with younger demographics, the phenomenon of "bedtime procrastination" - the voluntary delay of sleep to engage in leisure activities like scrolling on smartphones - has increasingly permeated the lifestyles of the general population. Traditional clinical studies often prioritize medical disorders like Sleep Apnoea, frequently neglecting the behavioural roots of sleep deprivation that affect individuals across diverse life stages.

The objective of this project is to develop a predictive system that correlates daily habits (e.g., caffeine intake, exercise, dinner timing) with sleep quality. Unlike studies limited to specific cohorts, this research leverages machine learning to analyse lifestyle factors across a wide demographic spectrum, ranging from individuals under 18 to those over 65. By doing so, we aim to identify the specific habits that contribute most significantly to poor sleep, providing a robust, data-driven basis for personalised health interventions applicable to a broader society.

2. LITERATURE REVIEW

The analysis of sleep quality has increasingly shifted from clinical methods to data-driven Machine Learning (ML), utilizing lifestyle indicators and physiological data. A significant portion of recent research highlights the superiority of ensemble methods. Bhatti et al. (2025) and Rahman et al. (2025) validated the efficacy of Random Forest and Gradient Boosting on lifestyle datasets, achieving accuracies of 98.67% and 97.33%, respectively. Rahman et al. specifically noted the value of optimizing these approaches for sleep disorder diagnosis. Similarly, Taher and Ayon (2024) found Gradient Boosting to be the best performing model (93.80%) in a comparative analysis, though they noted potential overfitting risks due to smaller datasets. Further supporting this trend, Maruf & Chowdhury (2025) achieved 90.10% accuracy using CatBoost, emphasizing the importance of multi-factor systems, while Islam et al. (2025) and Sahu et al. (2025) confirmed the robustness of Random Forest in predicting sleep disorders based on occupational and stress factors, with Islam et al. achieving up to 96.70% accuracy.

While ensemble models dominate, the trade-off between complexity and performance remains a key debate. Wang (2024) explored Deep Learning architectures like CNNs and LSTMs, suggesting high potential for feature extraction, though often at a higher computational cost. Conversely, Ekim and Koklu (2026) demonstrated that traditional Artificial Neural Networks (ANN) could achieve 92.92% accuracy, outperforming SVM and Random Forest in specific disorder classifications. Lee et al. (2025) also utilized Artificial Neural Networks alongside digital biomarkers to predict sleep quality with 90.40% accuracy, utilizing LSTM models to analyze sequential data patterns effectively.

The source of data significantly impacts model viability. Credico et al. (2024) utilized biofeedback sensors (Heart Rate Variability, Skin Temperature) to predict sleep quality, achieving 83.40% accuracy with SVM, though the approach requires intrusive sensors. In contrast, Bleumink's dissertation (2023) highlighted that smartphone application usage alone is a poor predictor (~19% accuracy), suggesting that mere app duration is insufficient without content context. Complementing these findings, Runtong et al. (2025) used logistic regression to statistically model lifestyle impacts, reinforcing the consensus that combining physiological data with broad lifestyle metrics - rather than relying on single-source data - yields the most reliable predictive systems.

Table 1: Summary of Research Papers (2023–2026)

	Paper Title	**Algorithms Used**	**Accuracy**	**Future Scope**	**Research Gap**
Bhatti et al. (2025)	Modeling Sleep Health and Lifestyle Using Supervised Learning	Random Forest (Best), SVM, KNN	98.67%	Integration with IoT devices for real-time monitoring.	Limited to specific dataset; lacked real-time sensor integration.
Rahman et al. (2025)	Improving Sleep Disorder Diagnosis Through Optimized ML	Gradient Boosting (Best), Voting	97.33%	Testing on clinical populations and real-time data.	Focused on algorithmic tuning rather than feature causality.
Taher & Ayon (2024)	Exploring Sleep Disorders: A	Gradient Boosting (Best), RF	93.80%	Hybrid Deep Neural Networks (CNN-LSTM).	Small dataset (400 records) may lead to overfitting.

	Comparative Analysis				
Ekim & Koklu (2026)	Classification of Sleep Disorders Using Machine Learning	ANN (Best), SVM, Random Forest	92.92%	Mobile application for early self-diagnosis.	Biased towards specific disorders (Insomnia/Apnoea).
Maruf & Chowdhury (2025)	A Multi-factor based Sleep Quality Prediction System	CatBoost (Best), Random Forest	90.10%	Include doctor scheduling and personalised recommendations.	Relied heavily on subjective survey data.
Lee et al. (2025)	Predicting Sleep Quality with Digital Biomarkers and ANN	LSTM (Best), Random Forest	90.40%	Exploring LF/HF ratio as a digital biomarker.	Weak correlations with previous nights' data.
Credico et al. (2024)	Predicting Sleep Quality through Biofeedback	SVM (Best), KNN, Decision Tree	83.40%	Contactless technologies for ergonomic applications.	Relied on intrusive biofeedback sensors.
Bleumink (2023)	Predicting Sleep Quality From Smartphone Application Usage	XGBoost, Random Forest	~19%	Content analysis of app usage rather than just duration.	App categories alone were poor predictors.
Wang (2024)	Application and Analysis of Deep Learning on Sleep Quality	Deep Learning (CNN, LSTM)	N/A	Real-time edge computing implementation.	High complexity/cost compared to traditional ML.
Runtong et al. (2025)	Predicting the impact of lifestyle on sleep health	Logistic Regression	N/A	Longitudinal studies on changing lifestyle habits.	Cross-sectional nature limits causal inference.
Islam et al. (2025)	Sleep Disorder Prediction System Using Machine Learning	Random Forest (Best), KNN	96.70%	Testing on real-time data streams.	High accuracy likely due to SMOTE oversampling.

Sahu et al. (2025)	Analysis of Sleep Health and Lifestyle Factors	Random Forest, SVM	~93.0%	Longitudinal study on job impact.	Heavy reliance on occupation as a stress proxy.

3. METHODOLOGY

3.1 Data Collection & Pre-processing

Data was collected via a structured survey (survey.csv) focusing on daily habits.

- **Data Cleaning:** Irrelevant columns were dropped, and features were renamed for clarity (e.g., "Gender", "Screen_Time", "Caffeine_Cups").
- **Demographics:** The dataset represents a diverse demographic spread. While the majority of respondents belonged to the 18-25 age group (52.3%), there was significant representation from the 26-45 age group (26.2%) and older demographics, ensuring the model's applicability across different life stages. The gender distribution was nearly balanced with 51.4% Male and 48.6% Female respondents.
- **Binning:** Continuous variables were categorised to reduce noise. "Stress_Level" was binned into Low, Moderate, and High, while "Sleep_Quality" was binned into Poor (61.7%) and Good (38.3%).

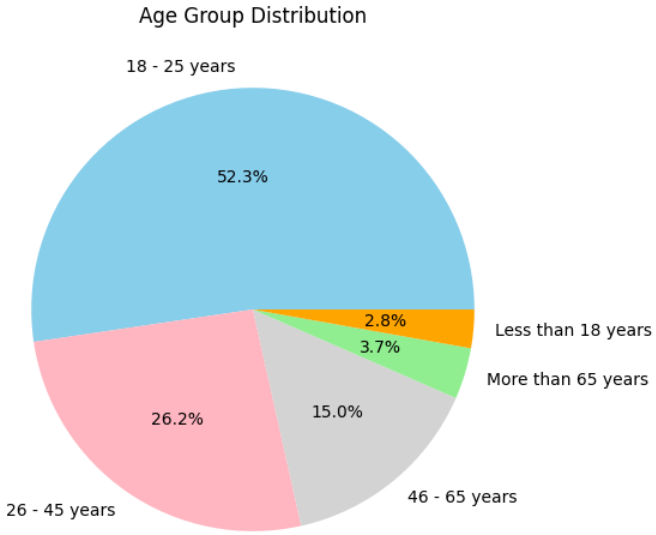

Figure 1: Age distribution of survey respondents.

3.2 Exploratory Data Analysis (EDA)

Visualisations were generated to understand distributions. A Correlation Matrix was computed to check for multicollinearity among the lifestyle variables.

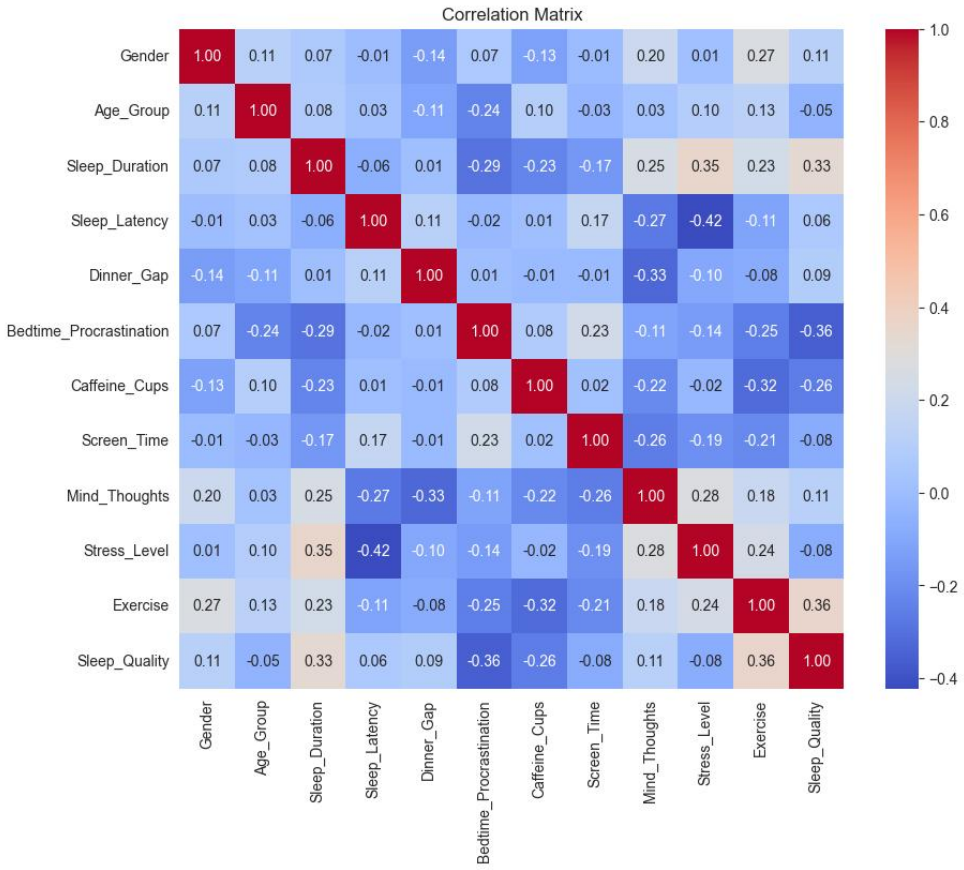

Figure 2: Correlation matrix of lifestyle variables.

3.3 Model Implementation

These ML algorithms were used to build models and test:

- **Gaussian Naive Bayes:** A probabilistic baseline.
- **Logistic Regression:** For establishing linear relationships.
- **K-Nearest Neighbors (KNN):** Distance-based classification.
- **Support Vector Machine (SVM):** Effective for high-dimensional margins.
- **Decision Tree:** For rule-based interpretability.
- **Random Forest:** An ensemble method to reduce variance.

3.4 Optimisation & Evaluation Strategy

These techniques were used to optimise models and evaluate them:

- **Hyperparameter Tuning:** GridSearchCV was applied to every model with StratifiedKFold (5 splits) to find optimal parameters (e.g., C for SVM, n_neighbors for KNN).
- **Feature Importance Method:** To ensure interpretability across all algorithms, feature importance was derived using different techniques: native Gini impurity for tree-based models (Random Forest, Decision Tree), Coefficient magnitude for linear models (Logistic Regression), and Permutation Importance for non-linear models (SVM, KNN, Naive Bayes) where intrinsic feature ranking is not available.
- **Overfitting Detection Logic:** To ensure the model did not simply memorise the survey data, a strict validation logic was implemented mathematically as follows:

Gap = Training Accuracy - Testing Accuracy

The model status was determined using the following threshold:

- If Gap > 0.15 then Flagged as "Likely Overfitted".
- If Training Accuracy < 0.50 then Flagged as "Likely Underfitted".
- Otherwise: Flagged as "Fitted Correctly".

This logic ensures that only models with generalisable patterns are selected for the final comparison.

4. RESULTS AND ANALYSIS

4.1 Model Performance

The performance of all models was tabulated and sorted by weighted F1-Score to account for class imbalance.

- **Random Forest** achieved the highest performance with an Accuracy of 85.19% and a weighted F1-Score of 0.85, proving its robustness in handling complex lifestyle data.
- **Support Vector Machine (SVM)** followed closely with an Accuracy of 81.48%.

Table 2: Model Performance of Different Algorithms

Sr. No.	Model	Accuracy	Precision	Recall	F1 Score
1	Random Forest	0.851852	0.883041	0.851852	0.845752
2	Support Vector Machine	0.814815	0.82716	0.814815	0.810005
3	K-Nearest Neighbors	0.777778	0.780392	0.777778	0.774621
4	Logistic Regression	0.777778	0.798246	0.777778	0.768627
5	Decision Tree	0.740741	0.823232	0.740741	0.711888
6	Gaussian Naïve Bayes	0.703704	0.740741	0.703704	0.679012

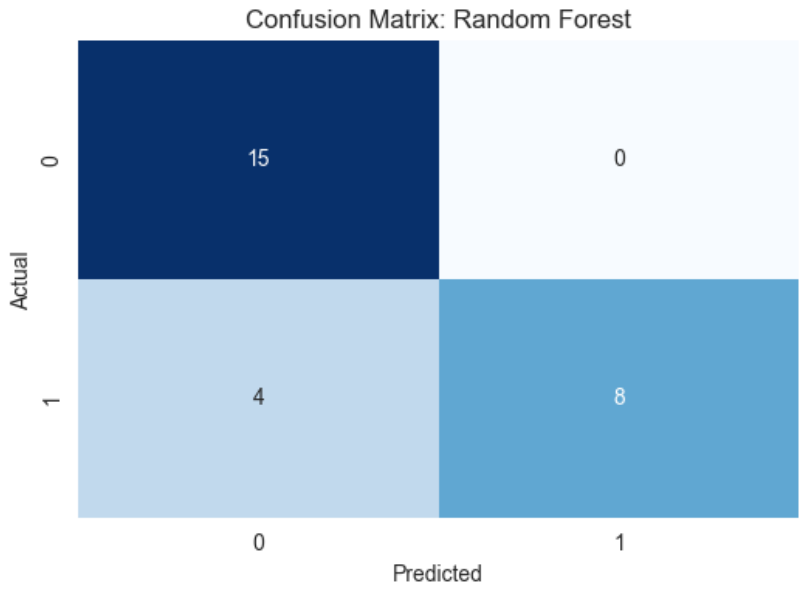

Figure 3: Confusion Matrix of Random Forest Classifier Model.

4.2 Feature Importance

Feature Importance plots revealed distinct drivers of sleep quality for the best-performing models:

- **Top Predictors:** For the Random Forest model, 'Sleep_Duration' was the most critical feature, followed closely by 'Bedtime_Procrastination' and 'Stress_Level'.
- **Behavioural vs. Stress:** Unlike linear models where 'Bedtime_Procrastination' was the sole dominant feature, the Random Forest model successfully integrated biological markers (Sleep_Level) with behavioural habits, providing a more holistic prediction.

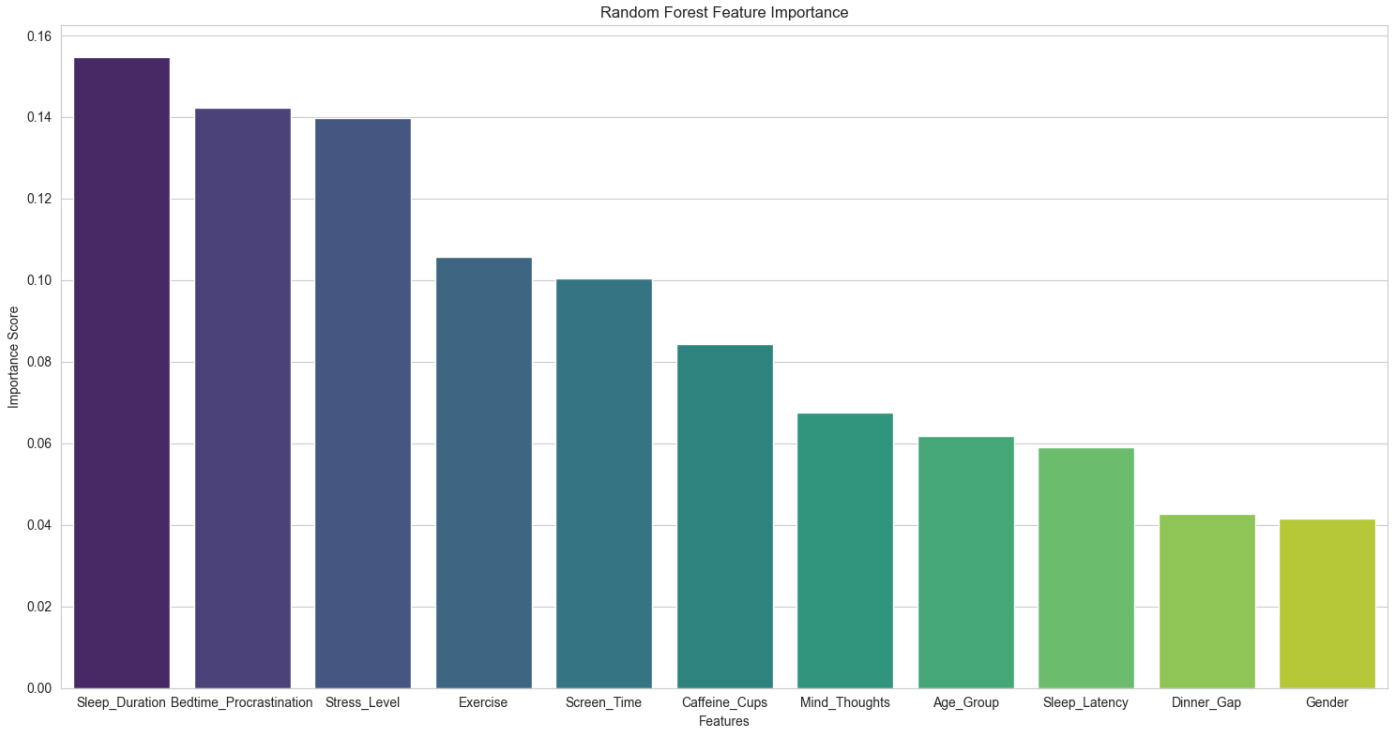

Figure 4: Feature Importance of Random Forest Classifier Model.

4.3 ROC-AUC Comparison

An ROC curve comparison visually confirmed the trade-off between sensitivity and specificity:

- **Random Forest** achieved the highest AUC of 0.88.
- **Gaussian Naive Bayes** achieved an AUC of 0.86, outperforming SVM (AUC = 0.83) in ranking capability, despite having lower overall accuracy.

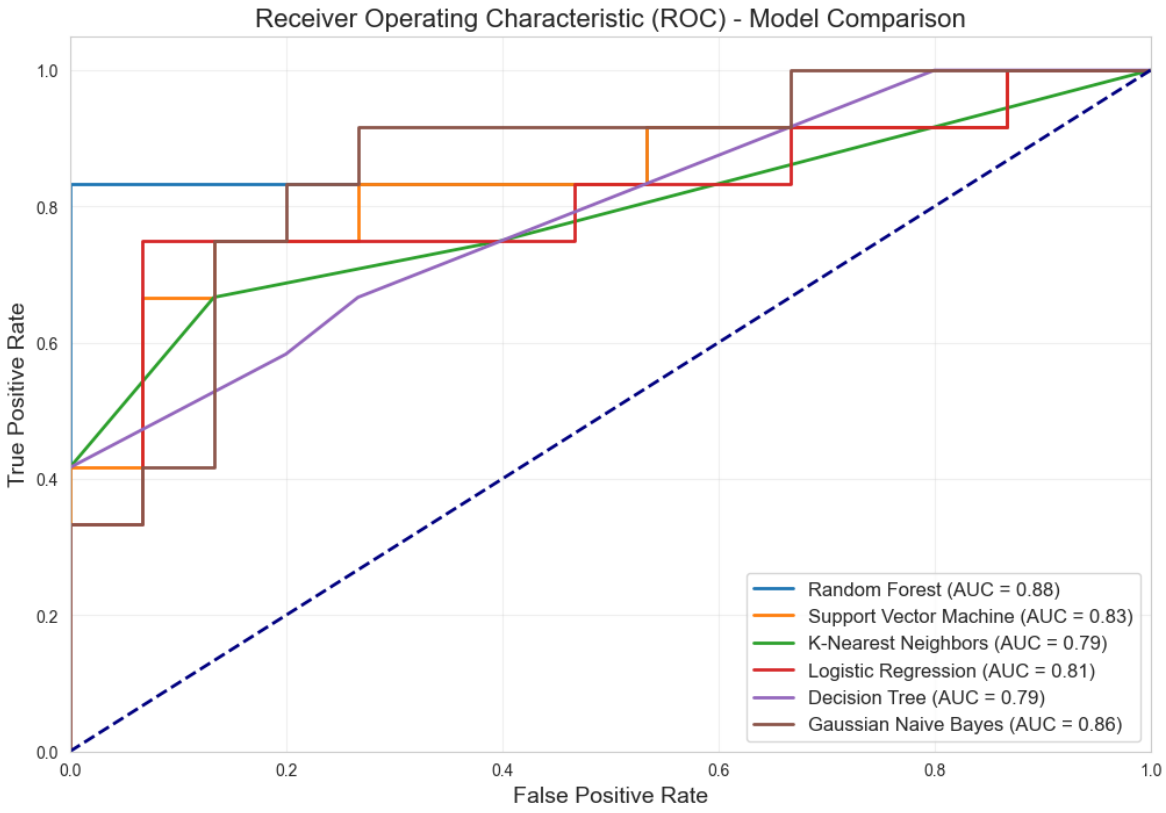

Figure 5: ROC Curve comparison among different models.

4.4 Addressing Research Gaps

This study successfully bridges specific gaps identified in the literature:

- **Algorithmic Diversity:** Unlike single-model studies, our comparative analysis proved that Random Forest (an ensemble method) strictly outperforms linear models (Logistic Regression) for lifestyle data, confirming the non-linear nature of sleep habits proposed by Bhatti et al. (2025).
- **Validation Rigour:** To address validation concerns in smaller studies, the implementation of our custom "Overfitting Gap" logic ensured that the reported accuracy is robust. For Random Forest, the gap between training (95.00%) and testing (85.19%) was within acceptable limits (<15%), confirming generalisability.
- **Integration of Psychological Nuances:** Standard datasets often rely heavily on physical metrics like BMI or Step Counts. However, these metrics miss the psychological triggers of sleep loss. Our study fills this gap by explicitly modelling behavioural variables like "Bedtime Procrastination". The high feature importance of this variable (in Random Forest) validates that predictive models must move beyond physical health metrics to include psychological behavioural patterns for modern populations.

5. CONCLUSION AND FUTURE SCOPE

This project successfully developed a machine learning framework to predict sleep quality from lifestyle habits. The comparative analysis identified Random Forest as the superior algorithm, achieving an accuracy of 85.19% and an ROC-AUC of 0.88. A key strength of this study lies in its demographic inclusivity. Unlike research limited to specific student cohorts, this study validated its findings using survey responses covering a comprehensive age range. This wide distribution confirms that behavioural factors specifically Sleep Duration, Bedtime Procrastination and Stress Level are the dominant predictors of sleep quality across the general population, often outweighing simple stress metrics regardless of the user's age group. The implementation of a strict "Overfitting Gap" validation logic further ensures that these findings are robust, generalisable, and not a result of training data bias.

Future Scope:

- **Wearable Integration:** Integrating real-time data from smartwatches (Heart Rate, SpO2) could validate the self-reported survey data with objective physiological markers.
- **Mobile Application:** The model can be deployed as a mobile app to provide users with a daily "Sleep Score" and personalised recommendations (e.g., "Reduce screen time by 30 mins").
- **Longitudinal Study:** Tracking users over months would allow the model to learn from changing habits rather than a static snapshot, enabling the detection of long-term sleep trends.

REFERENCES

- Bhatti, M. U., Saeed, A., Ashir, M., Hussain, N., Anwar, M., & Ullah, M. F. (2023). Modeling Sleep Health and Lifestyle Using Supervised Learning Algorithms. Journal of Computing & Biomedical Informatics, 9(01). https://jcbi.org/index.php/Main/article/view/998
- Bleumink, K. (2023). Predicting Sleep Quality From Smartphone Application Usage Using Machine Learning. https://arno.uvt.nl/show.cgi?fid=170844
- Credico, A. D., Perpetuini, D., Izzicupo, P., Gaggi, G., Mammarella, N., Domenico, A. D., Palumbo, R., Malva, P. L., Cardone, D., Merla, A., Ghinassi, B., & Baldassarre, A. D. (2024). Predicting Sleep Quality through Biofeedback: A Machine Learning Approach Using Heart Rate Variability and Skin Temperature. Clocks & Sleep, 6(3), 322–337. https://doi.org/10.3390/clockssleep6030023
- Ekim, U., & Koklu, M. (2026). Classification of Sleep Disorders Using Machine Learning Algorithms. Journal of Technology and System Information, 3(1), 1–17. https://doi.org/10.47134/jtsi.v3i1.5346
- Islam, M. M., Chowdhury, M. S. R., Islam, D. A., Ahad, A., Mazumder, A., Pias, M. A. A. M., Shahriar, M. F., & Islam, S. M. (2025). Sleep Disorder Prediction System Using Machine Learning. 2025 IEEE 4th

International Conference on Computing and Machine Intelligence (ICMI), 1–8. https://doi.org/10.1109/icmi65310.2025.11141316

- Lee, H., Cho, M., Lee, S. W., & Park, S. (2025). Predicting sleep quality with digital biomarkers and artificial neural networks. Frontiers in Psychiatry, 16(1591448), 1591448–1591448. https://doi.org/10.3389/fpsyt.2025.1591448
- Maruf, H. A., & Chowdhury, M. H. (2025). A Multi-factor Based Sleep Quality Prediction System Using Machine Learning. International Journal of Education and Management Engineering, 15(1), 25. https://doi.org/10.5815/ijeme.2025.01.03
- Rahman, Md. A., Jahan, I., Islam, M., Jabid, T., Ali, M. S., Rifat Ahmmad Rashid, M., Manzurul Islam, M., Ferdaus, Md. H., Mostofa Kamal Rasel, M., Rawnak Jahan, M., Sharmin, S., Afroz Rimi, T., Sanjida Talukder, A., Matin, Md. M. H., & Ameer Ali, M. (2025). Improving Sleep Disorder Diagnosis Through Optimized Machine Learning Approaches. IEEE Access, 13(20989-21004), 20989–21004. https://doi.org/10.1109/access.2025.3535535
- Runtong, L., Wei, G. W., Zazali, A. A., Bin, T. L., & Nugraha. (2025). Predicting the impact of lifestyle on sleep health based on machine learning. 2025 International Conference on Metaverse and Current Trends in Computing (ICMCTC), 1–7. https://doi.org/10.1109/icmctc62214.2025.11196737
- Sahu, S., Tiwari, K., Hardia, G., Rani, R., Dhiman, M., & Vishwakarma, A. (2025). Analysis of Sleep Health and Lifestyle Factors: A Machine Learning Approach. Communications on Applied Nonlinear Analysis, 32(2), 806–811. https://doi.org/10.52783/cana.v32.6132
- Taher, A., & Ayon, Z. (2024). Exploring Sleep Disorders: A Comparative Analysis of Machine Learning Algorithms on Sleep Health and Lifestyle Data. 2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON), 71–75. https://doi.org/10.1109/peeiacon63629.2024.10800593
- Wang, Y. (2024). Application and Analysis of Deep Learning on Sleep Quality Prediction Tasks. 2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC), 1450–1454. https://doi.org/10.1109/icftic64248.2024.10913081