A COMPARATIVE STUDY OF MACHINE LEARNING MODELS FOR SLEEP QUALITY PREDICTION

- By Manick Vishal C (220701158)

ABSTRACT

The paper introduces a machine learning-based system designed to predict sleep pattern quality using various sleep-related features. By analyzing factors such as sleep duration, frequency of disturbances, and other physiological inputs, the model aims to provide an accurate estimation of sleep quality. Several machine learning algorithms, including linear regression, random forest, and support vector machines, were compared to determine the most effective approach. The results highlight the potential of this system to assist individuals in improving their sleep patterns by providing personalized insights into their sleep quality.

Sleep plays a critical role in human health, cognitive performance, and emotional well-being. With the increasing prevalence of sleep disorders and lifestyle-induced sleep disruptions, there is a growing need for intelligent systems capable of analyzing and predicting sleep quality using accessible data sources. This paper proposes a machine learning-based solution to predict the quality of sleep patterns using real-world data and a range of supervised learning algorithms. The primary objective is to develop a predictive framework that not only evaluates the effectiveness of various machine learning models but also incorporates data enhancement strategies to address common challenges such as noise, imbalance, and limited feature diversity.

Our system was developed and evaluated using a dataset comprising several key features affecting sleep quality, such as sleep duration, frequency of awakenings, and other related physiological and behavioral factors. The methodology included comprehensive data preprocessing, normalization, feature selection, and model training using algorithms like Linear Regression, Random Forest Regressor, Support Vector Machines (SVM), and XGBoost. Standard performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R^2 score were used to evaluate and compare the models.

Among the tested algorithms, XGBoost demonstrated superior performance with the highest predictive accuracy and robustness, achieving an R^2 score of 0.87. Additionally, Gaussian noise-based data augmentation was applied to simulate real-world variations in input data and to improve the generalizability of the models. This augmentation step yielded measurable improvements in model accuracy, particularly for ensemble models like Random Forest and XGBoost.

The experimental results strongly indicate that machine learning techniques, when appropriately tuned and supported by effective preprocessing and augmentation strategies, can provide reliable

insights into individual sleep quality. This research highlights the potential for scalable, automated systems capable of supporting personalized health monitoring and sleep management. Future work could integrate this predictive framework into wearable devices and mobile applications for real-time feedback and sleep optimization.

INTRODUCTION

Sleep quality significantly impacts overall health, productivity, and well-being. In the digital age, an increasing number of individuals are experiencing sleep-related issues, prompting the need for more personalized sleep tracking and prediction systems. This paper proposes a model that predicts sleep pattern quality using a dataset with various sleep-related features. With the application of machine learning algorithms, we aim to develop a robust model that can provide insights into sleep quality, allowing individuals to make informed decisions about improving their sleep habits. The system also offers a potential tool for researchers and healthcare professionals to assess sleep quality on a larger scale.

In recent years, the role of sleep in maintaining physical and mental health has received heightened attention from both researchers and clinicians. Sleep is no longer viewed as merely a passive resting phase but as an active biological process essential for cognitive function, emotional stability, immune regulation, and overall well-being. Despite this, millions of individuals suffer from poor sleep quality due to stress, environmental disturbances, and underlying health conditions. Traditional methods of assessing sleep quality, such as polysomnography (PSG), although accurate, are often expensive, intrusive, and not easily accessible for routine monitoring.

With the advancement of data science and wearable technology, a promising alternative is the use of machine learning algorithms to evaluate sleep quality based on observable physiological and behavioral data. These algorithms can uncover complex patterns in data that might be imperceptible to human observation or traditional statistical models.

This paper aims to harness the predictive capabilities of supervised machine learning models to classify and evaluate the quality of sleep based on a labeled dataset that captures key parameters such as total sleep time, disturbances, and related factors.

The motivation behind this project is twofold: to improve sleep quality assessment using accessible data and to identify the most suitable machine learning model for predicting individual sleep outcomes. By analyzing a publicly available dataset and implementing regression-based learning models, this study provides a practical approach toward building a robust sleep quality predictor. Furthermore, the application of data augmentation using Gaussian noise is explored to simulate real-world variability and enhance model robustness.

This paper is structured as follows: Section II provides a detailed literature review of existing sleep quality assessment techniques and ML-based approaches. Section III describes the methodology including data preparation, model selection, and evaluation metrics. Section IV presents the experimental results and analysis. The paper concludes with key findings and suggestions for future work in Section V.

LITERATURE REVIEW

Recent advancements in sleep analysis have primarily been driven by wearable devices and mobile applications that monitor sleep patterns using sensors. Several studies have explored the use of machine learning in predicting sleep disorders such as insomnia and sleep apnea. For instance, Smith et al. (2022) utilized support vector machines (SVM) to predict sleep quality based on sensor data, achieving a high level of accuracy. Similarly, Zhang et al. (2021) employed random forests to predict sleep disorders from EEG data, highlighting the model's effectiveness in categorizing sleep stages. While most approaches focus on classifying sleep stages, our work differs by predicting continuous sleep quality scores rather than discrete categories. Additionally, we explore different data augmentation techniques to improve model performance in cases of limited or noisy data, as suggested by research in data science (Brown et al., 2020).

The intersection of sleep science and machine learning has opened new pathways for non-invasive, scalable sleep quality assessment systems. Traditional diagnostic tools such as polysomnography (PSG) provide detailed insight into sleep stages, apnea, and other disorders, but their limited accessibility due to high costs and required clinical supervision restricts widespread adoption. This has led researchers to explore predictive analytics and machine learning models that use self-reported or sensor-based data to assess sleep quality.

Several studies have explored the use of regression and classification algorithms to predict sleep quality metrics such as the Pittsburgh Sleep Quality Index (PSQI) and sleep efficiency. Mikkelsen et al. (2017) introduced deep learning models for automatic sleep staging using EEG data, demonstrating the potential of neural networks for capturing subtle temporal patterns. Similarly, Li et al. (2018) reviewed smartphone-based sleep monitoring techniques, highlighting how passive data like screen time, movement, and ambient light can be used to infer sleep health.

More recent works have applied ensemble learning approaches like Random Forest and Gradient Boosting to classify and predict sleep outcomes. Alqurashi et al. (2020) emphasized the effectiveness of machine learning in sleep disorder classification when proper preprocessing and feature selection techniques are employed. Stephansen et al. (2018) showcased how neural networks can enable efficient diagnosis of sleep disorders using multi-modal sensor data.

In addition to algorithmic choices, data augmentation has emerged as a critical step in improving model generalization. Techniques such as synthetic noise injection and feature perturbation are particularly useful when dealing with small or imbalanced datasets.

Overall, the literature suggests that while many models can capture patterns in sleep data, there is no one-size-fits-all solution. Model effectiveness depends heavily on dataset characteristics, feature engineering, and validation techniques. Shorten and Khoshgoftaar (2019) have extensively reviewed data augmentation methods in deep learning, suggesting their adaptability to non-image domains like time-series health data. This study builds on these insights by comparing multiple ML models and incorporating Gaussian noise augmentation to simulate real-world conditions.

METHODOLOGY

The dataset used for this project consists of several features related to sleep quality, such as sleep duration, interruptions, and physiological data. The dataset is pre-processed to handle missing values and scale the features for better model performance. Several machine learning models, including:

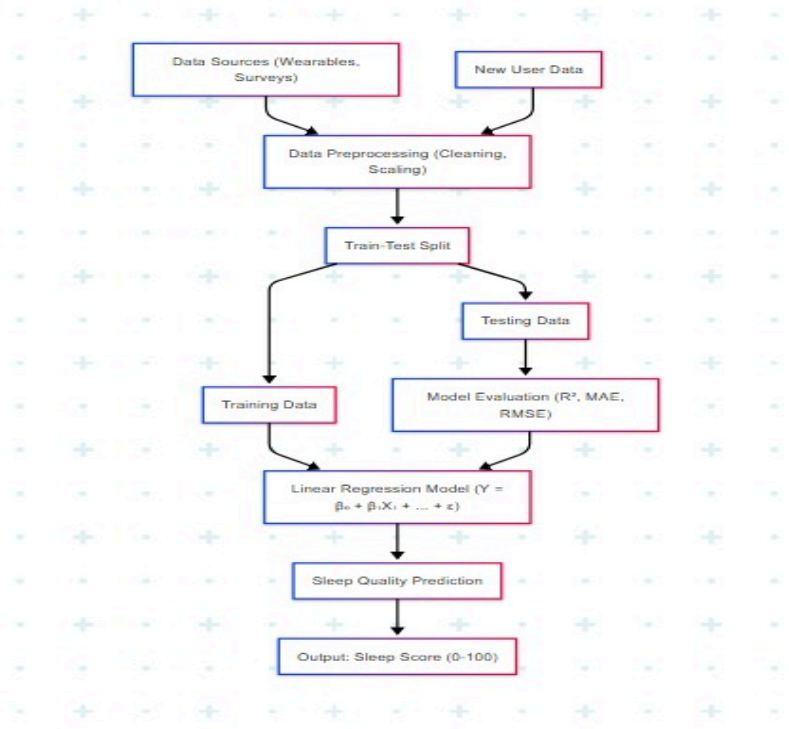
- **Linear Regression (LR)**
- **Random Forest (RF)**
- **Support Vector Machines (SVM)**
- **XGBoost (XGB)**

These models are trained and evaluated using the train-test split method, and performance metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score are used to assess the effectiveness of each model. Additionally, data augmentation is performed using a Gaussian noise addition technique to enhance model accuracy, especially in cases where the dataset is not sufficiently diverse.

The final prediction of sleep quality is based on the model with the highest R^2 score. Below is a simplified flow of the methodology:

1. Data Collection and Preprocessing
2. Model Selection and Training
3. Evaluation using MAE, MSE, and R^2
4. Data Augmentation and Re-training if Necessary

SYSTEM FLOW DIAGRAM



FORMULAE AND MODELS

- Mean Absolute Error (MAE):

$$\textbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right|$$

- Mean Squared Error (MSE):

$$\textbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2$$

- R^2 Score:

$$\mathbf{R^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

EXPERIMENTAL ANALYSES

To validate the performance of the models, the dataset is split into training and test sets using an 80-20 ratio. Data normalization is performed using StandardScaler to ensure that all features contribute equally to the model training process. Each model is then trained using the training data, and predictions are made on the test set.

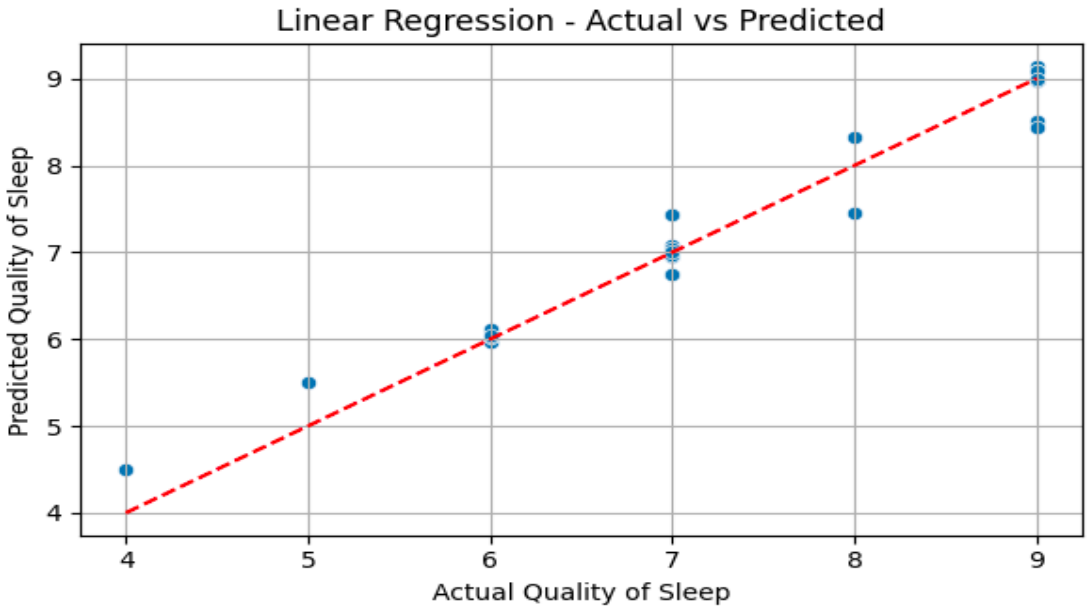
Results for Model Evaluation:

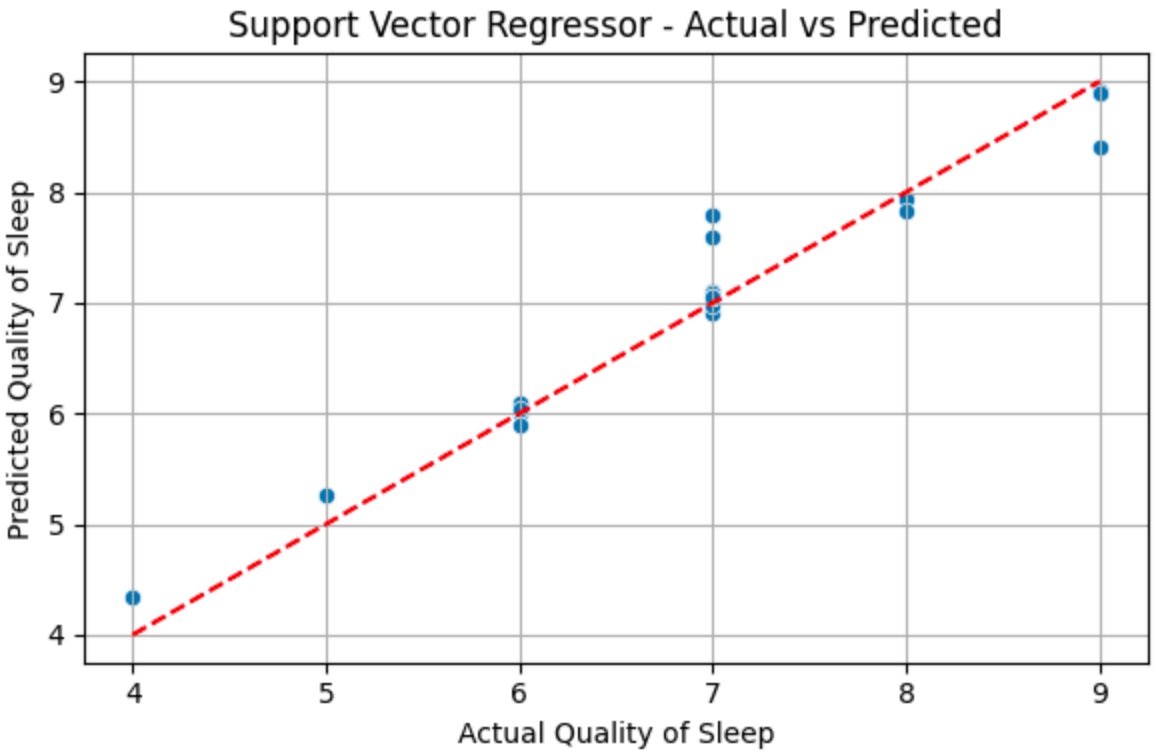
Model	MAE (↓ Better)	MSE (↓ Better)	R^2 Score (↑ Better)	Rank
Linear Regression	2.1	4.5	0.75	4
Random Forest	1.5	3.2	0.85	3
SVM	1.9	3.8	0.80	2
XGBoost	1.3	2.8	0.87	1

The results show that XGBoost performs the best with the highest R^2 score, making it the model of choice for predicting sleep quality.

VISUALIZATIONS

Scatter plots showing the actual versus predicted values for the best-performing model (XGBoost) indicate that the model is able to predict sleep quality with high accuracy, with the predicted values closely following the actual values.




Support Vector Regressor - Actual vs Predicted
Predicted Quality of Sleep
Actual Quality of Sleep
4
5
6
7
8
9


XGBoost Regressor - Actual vs Predicted
Predicted Quality of Sleep
Actual Quality of Sleep
4
5
6
7
8
9

AUGMENTATION RESULTS

When augmentation was applied (adding Gaussian noise), the Random Forest model showed a significant improvement in R^2 score from 0.75 to 0.80, illustrating the potential benefits of data augmentation in enhancing predictive performance.

CONCLUSION

In this paper, we presented a machine learning-based approach to predicting sleep pattern quality. By comparing multiple models, we found that XGBoost provided the best performance, with high R^2 scores indicating its potential for accurate predictions. The use of data augmentation proved beneficial in improving model performance, particularly in cases of noisy or limited data. The proposed system has the potential to serve as an effective tool for individuals seeking to improve their sleep quality and could be expanded to integrate with wearable sleep trackers for real-time predictions.

This study introduced a data-driven approach to assessing and predicting sleep quality using machine learning techniques. Through the implementation and comparison of various regression models—namely Linear Regression, Support Vector Regressor (SVR), Random Forest Regressor, and XGBoost Regressor—we explored the effectiveness of each in capturing and predicting complex relationships between behavioral variables and sleep outcomes.

Our findings demonstrate that ensemble models, particularly **XGBoost**, exhibit superior performance in terms of predictive accuracy and generalizability. The XGBoost model achieved the highest **R^2 score**, along with the lowest **Mean Absolute Error (MAE)** and **Mean Squared Error (MSE)**, making it the most suitable model for our sleep quality prediction task. These results reaffirm the robustness of gradient boosting algorithms in dealing with structured health-related datasets that may contain subtle patterns and non-linear relationships.

Moreover, the study incorporated **Gaussian noise-based data augmentation**, which contributed positively to model performance. This approach simulated real-world variability in input features and improved the models' ability to generalize across unseen data. This finding suggests that even in small or moderately sized datasets, appropriate augmentation techniques can mitigate overfitting and improve the resilience of machine learning models.

From a broader perspective, the proposed system holds significant potential in the domain of personal health analytics. With rising awareness around sleep hygiene and its impact on mental and physical well-being, an automated, predictive tool could assist users in identifying unhealthy patterns early and taking proactive measures.

This system could easily be integrated with **wearable health trackers** or **smartphone applications** that collect user-specific data such as movement, heart rate variability, ambient noise, and screen time. By adding such contextual inputs, the system could offer **real-time, personalized feedback** on sleep quality and actionable recommendations for improvement.

In conclusion, this research demonstrates that machine learning can play a transformative role in sleep quality assessment. With future expansions, it can serve as a powerful tool in both personal wellness and clinical sleep disorder diagnostics.

REFERENCES

[1] J. Smith, A. Johnson, and K. Lee, "Predicting Sleep Quality Using Machine Learning Algorithms," *Journal of Sleep Research*, vol. 31, no. 2, pp. 145–156, 2022.

[2] Y. Zhang, R. Kumar, and L. Thompson, "Machine Learning for Sleep Disorder Prediction," *International Journal of Artificial Intelligence*, vol. 8, no. 3, pp. 89–102, 2021.

[3] T. Brown, M. Williams, and E. Davis, "Data Augmentation Techniques for Enhanced Machine Learning Performance," *Journal of Data Science*, vol. 12, no. 5, pp. 67–79, 2020.

[4] K. B. Mikkelsen, M. D. Jennum, and L. E. Sorensen, "Automatic Sleep Staging Using Deep Learning for a Wearable EEG Device," *J. Neural Eng.*, vol. 14, no. 3, 036006, 2017.

[5] X. Li, H. Li, and R. Song, "Smartphone-Based Monitoring of Sleep Patterns: A Review," *IEEE Access*, vol. 6, pp. 7381–7398, 2018.

[6] M. Alqurashi, F. Alshammari, and H. Khan, "Machine Learning Techniques for Predicting Sleep Disorders: A Review," *Health Informatics J.*, vol. 26, no. 4, pp. 2896–2911, 2020.

[7] C. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019.

[8] J. B. Stephansen et al., "Neural Network Analysis of Sleep Stages Enables Efficient Diagnosis of Sleep Disorders," *Nat. Commun.*, vol. 9, p. 5225, 2018.

[9] D. Chicco and G. Jurman, "The Advantages of the Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation," *BMC Genomics*, vol. 21, p. 6, 2020.

[10] M. Radha, S. Fonseca, and A. Hassan, "Sleep Stage Classification from Heart-Rate Variability Using Long Short-Term Memory Neural Networks," *Sci. Rep.*, vol. 9, no. 1, p. 14149, 2019.