Comparative Analysis of Machine Learning Models for Sleep Data Analysis: Evaluating R2 Scores as Performance Metrics

Wilson Anthony Widjaja
Department of Computer Science
Bina Nusantara University
Jakarta, Indonesia
wilson.anthony@binus.ac.id

Abyaan Syauqi Muhammad
Department of Computer Science
Bina Nusantara University
Jakarta, Indonesia
abyaan.muhammad@binus.ac.id

Andre Hanjaya
Department of Computer Science
Bina Nusantara University
Jakarta, Indonesia
andre.hanjaya001@binus.ac.id

Yohan Muliono
Department of Cyber Security
Bina Nusantara University
Jakarta, Indonesia
ymuliono@binus.edu

Simeon Yuda Prasetyo
Department of Computer Science
Bina Nusantara University
Jakarta, Indonesia
simeon.prasetyo@binus.edu

***Abstract*— Sleep data analysis plays a crucial role in understanding and addressing various sleep-related disorders, leading to improved overall health and well-being. With the advent of advanced technology and the increasing availability of wearable devices capable of recording sleep patterns, there is a growing need for robust machine learning models that can effectively analyze and interpret this data. The aim of this research is to perform a comparative analysis of machine learning models for sleep data analysis, specifically focusing on evaluating the R2 scores as performance metrics. Traditional sleep assessing methods have relied on subjective evaluations and limited sample sizes, which can introduce biases and reduce the generalizability of the findings. Moreover, the manual interpretation of sleep data is time-consuming and prone to human error. We can use machine learning for better understanding sleep patterns and developing personalized interventions. Therefore, selecting an appropriate machine learning model is essential for accurate prediction of sleep-related outcomes. This study presents a comparative analysis of multiple machine learning models for sleep data analysis, focusing on the evaluation of training and test R-squared scores as performance metrics. The choice of the best model depends on specific project requirements, considering factors such as model complexity, interpretability, computational demands, and domain knowledge. These findings provide valuable insights for researchers, clinicians, and sleep experts in selecting suitable machine learning models for sleep data analysis, paving the way for improved sleep research and personalized interventions. Future research can address identified limitations, such as dataset expansion and exploring ensemble approaches, to further enhance the accuracy and generalization of sleep data analysis models.**

Keywords*— *sleep quality, sleep, r-squared score, machine learning models, comparative analysis

I. INTRODUCTION

Sleep plays a crucial role in maintaining overall health and well-being, influencing various cognitive, physiological, and psychological processes. Evaluating the quality of sleep is crucial for understanding its impact on various aspects of a person's life, including learning, physical abilities, and performance. Unfortunately, in today's fast-paced world, a significant number of individuals experience poor sleep quality. Consequently, there is a growing interest in developing automated methods to measure sleep quality, which can aid in evaluating the effectiveness of treatments for common sleep disorders like restless legs syndrome, insomnia, narcolepsy, and obstructive sleep apnea. Analyzing sleep data is essential for understanding sleep patterns, diagnosing sleep disorders, and developing personalized interventions to improve sleep quality [1]. With the growing availability of sleep data collected through wearable devices [2] and other monitoring technologies, machine learning techniques have emerged as valuable tools for analyzing and predicting sleep-related outcomes [3].

The goal of this study is to evaluate and compare the performance of different machine learning models in analyzing sleep data. Specifically, we focus on assessing the goodness of fit and generalization capabilities of these models using R-squared scores as performance metrics [4]. By identifying the most effective model for sleep data analysis, we aim to provide researchers, clinicians, and sleep experts with valuable insights and recommendations for selecting appropriate machine learning approaches.

A number of studies have investigated the use of machine learning techniques for the analysis of sleep data, aiming to enhance the accuracy and efficiency of sleep disorder detection. Here are some of the relevant works that we've found and discussed.

D. H. Maulud and A. M. Abdulazeez discusses various works by different researchers on linear regression and polynomial regression and compares their performance using the best approach to optimize prediction and precision. Almost all of the articles analyzed in this review is focused on datasets in order to determine a model's efficiency, it must be correlated with the actual values obtained for the explanatory variables [5].

D. Radulović and D. Negovanović uses Multi-Layer Perceptron Regressor or MLP Regressor to predict gait speed based on walking parameters. This study [6] concludes the performance of the regressor model depends mostly on the number of hidden layers of the neural network which define its structure.

Cai et al. paper [7], uses a sample-rebalanced and outlier-rejected k-nearest neighbor regression model for short-term traffic flow forecasting. The experimental results, evaluated

on four real-world benchmark datasets, demonstrate the superiority of the proposed K-NN regression model over other models in short-term traffic flow forecasting.

Cai et al. uses the gradient boosting regression algorithm [8] to predict the Net Ecosystem Carbon Exchange (NEE) based on meteorology and flux data. Gradient boosting is an ensemble learning technique that combines multiple weak prediction models to create a strong predictive model.

Dewi C. utilizes random forest (RF) as a robust algorithm for feature selection and support vector machine (SVM) [9] for classification and regression tasks. Dewi C. states that the combination of RF, SVM, and tuned SVM regression can improve the performance of the regression model.

Feature selection plays a crucial role in dealing with datasets that contain a large number of variables. One robust algorithm that has emerged in this context is random forest (RF) [10]. RF is known for its ability to handle feature selection problems effectively and efficiently, particularly in regression tasks.

Another study in [11] has proven that machine learning can learn and accurately predict sleep disorders such as rapid eye behavior disorder (RBD) from electroencephalographic data. They proposed a method using Random Forest Classifier to determine and predict RBD sick patients with an accuracy of 90%.

In a study conducted by Park et al. [12], a comparative analysis was done to evaluate the performance of three machine learning models: Logistic Regression, Random Forest, and 1D-CNN. The researchers aimed to identify factors that correlate with participants' sleep quality. By utilizing these correlated features, the algorithms achieved an accuracy of 79.2% for Logistic Regression, 83.1% for Random Forest, and 87.2% for 1D-CNN.

The evaluation of the machine learning models involves training them on a subset of the sleep data [13] and assessing their performance on unseen data through appropriate cross-validation techniques. The primary performance metric used is the R-squared score, which measures the proportion of variance in the sleep data explained by the models.

Through this comparative analysis, we aim to identify the models that achieve the highest R-squared scores, indicating a better fit to the sleep data and improved predictive capabilities [14]. We also examine the generalization performance of the models by assessing their R-squared scores on unseen data. Furthermore, we discuss the implications of the findings, considering factors such as model complexity, interpretability, computational requirements, and domain knowledge.

In summary, this study examines several popular machine learning models commonly used in sleep research and known for their versatility in handling diverse data and relationships. The models considered include Linear Regression, MLP Regressor, KNN Regressor, Gradient Boosting, SVM, and Random Forest.

The outcomes of this study will contribute to the field of sleep research by providing insights into the strengths and weaknesses of different machine learning models for sleep data analysis. The findings will assist researchers, clinicians, and sleep experts in making informed decisions regarding the selection of the most appropriate machine learning approach for their specific sleep-related studies and interventions.

In the subsequent sections, we present the methodology used, describe the dataset and feature extraction process, outline the machine learning models employed, discuss the evaluation metrics, present the results, and provide a comprehensive analysis of the findings. Finally, we offer recommendations for selecting the most suitable machine learning model for sleep data analysis and suggest potential avenues for future research.

II. METHODOLOGY

The proposed approach methodology involves several crucial steps:

2.1 Datasets

The data used in this research was taken from [15] which contains sleep duration, predicted sleep quality from Sleep Cycle iOS application from 2014 – 2022.

No	Start time in second	End time in second	Time in bed	Sleep quality
0	82669	27013	30744.0	100
1	76670	77634	964.0	3
2	81769	26011	30642.0	98
3	81061	21781	27120.0	65
4	79930	17795	24265.0	72

Table 1. Sleep data example after thoroughly examined and extracted based on the features

2.2 Data Pre-processing

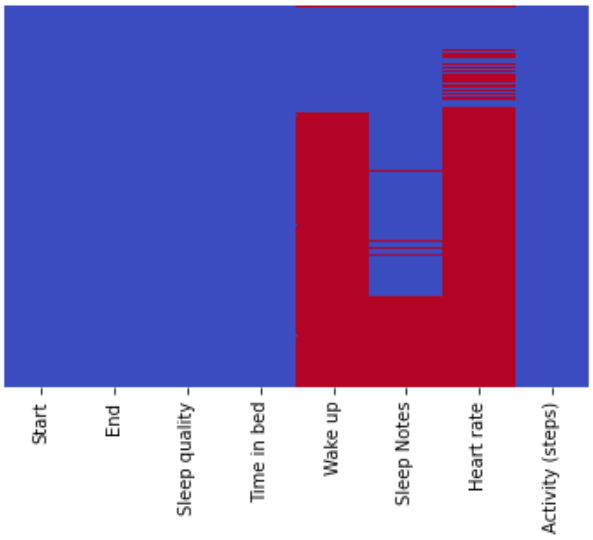

Fig 1. Plotting of the data (red part is missing data)

Firstly, the sleepdata.csv file is loaded into Python, using a provided python library "pandas". Figure 1 showed plotting of the dataset. After loading the dataset, it is observed that there are missing values within the data. Missing data can hinder the accuracy and reliability of machine learning models, as they require complete and consistent datasets to learn from. To address this issue, a common approach is to fill in the missing values with a suitable estimate.

In this case, the missing data is filled with the average mean of the datasets. This means that for each feature with missing values, the average value of that particular feature across the entire dataset is used to replace the missing entries. This imputation strategy helps to ensure that the data remains representative and preserves the overall distribution of the feature.

Once the missing values are filled, the next step involves scaling the data. Scaling is necessary to avoid biased learning in machine learning algorithms, especially those that are sensitive to the magnitude or range of the input features. Scaling ensures that all features have a similar range or distribution, preventing any one feature from dominating the learning process.

Common scaling techniques include standardization (also known as z-score normalization) or normalization (also known as min-max scaling). Standardization transforms the data to have a mean of zero and a standard deviation of one, while normalization scales the data to a specific range, often between 0 and 1.

By scaling the data, the machine learning model can effectively learn from each feature's contribution without being disproportionately influenced by features with larger values or wider ranges.

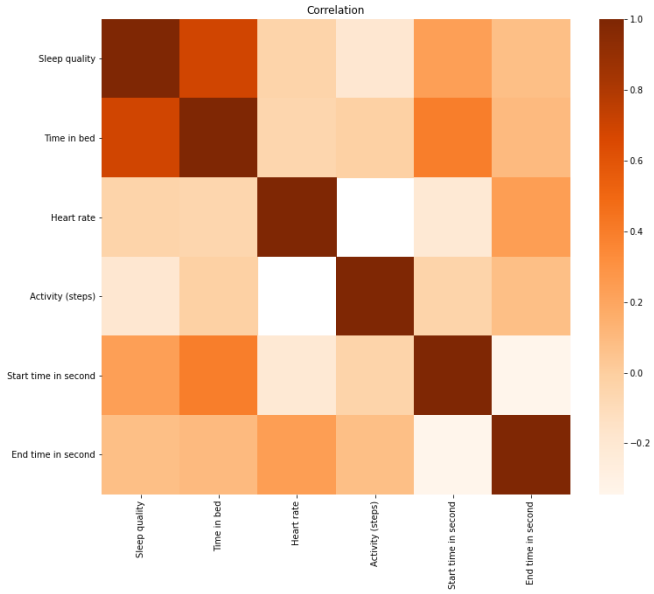

Fig 2. Correlation heatmap between features of the datasets

After conducting an in-depth analysis of the data and visualizing the relationships through a heatmap shown in Figure 2, we aimed to uncover the features that have the strongest correlation with the "sleep quality" variable. From the graphical representation, it became evident that two features stood out with a profound relationship: "sleep quality" and "time in bed". The heatmap showcased a distinct and prominent connection between these two variables, indicating that the duration of time spent in bed significantly impacts the quality of sleep.

On the other hand, the remaining features depicted a relatively weaker or insignificant association with each other and with "sleep quality". The heatmap revealed a lack of pronounced correlations among these variables, indicating that they have limited influence on sleep quality when compared to the crucial factor of "time in bed".

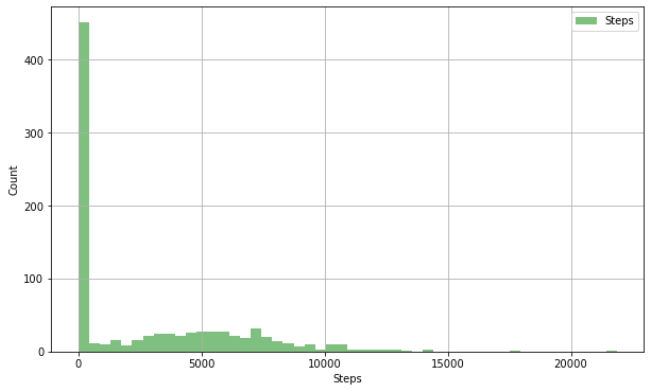

Fig. 3. Histogram of steps in the dataset taken

Further evaluating the dataset features, we discussed about the “steps” feature. From the figure 3 we plotted a chart to show how many steps the users take in the dataset. We found that most of the time steps are under 500. We also found activity in steps is when the user is awake. From our own experience, it is assumed that a large amount of activity can generate better sleep quality.

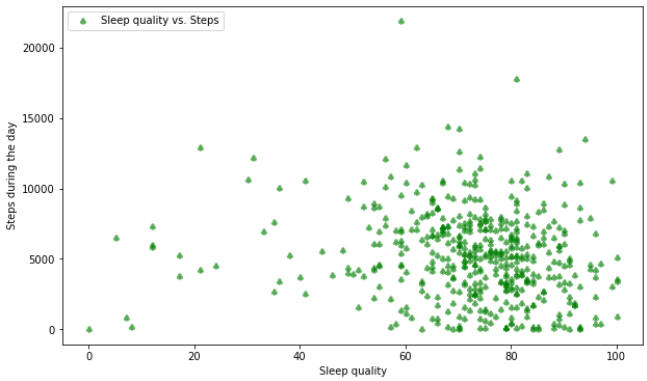

Fig . 4. Correlation between sleep quality and steps

To test our hypothesis in the above paragraph, we conducted an analysis to examine the correlations between different features. Surprisingly, our findings indicated that the features we investigated were not correlated. The plotted data in Figure 4 confirmed this observation, specifically when mapping out the relationship between "activity" and "sleep quality" features. The plot exhibited a scattered pattern, suggesting a lack of significant correlation between the two variables.

Based on this data, we arrived at the conclusion that the "activity" feature holds little to no significance in assessing sleep quality. Consequently, we have made the informed decision to exclude the "activity" feature from our analysis. By doing so, we aim to streamline our investigation, focusing on the features that possess more meaningful correlations with sleep quality, and thereby ensuring the accuracy and reliability of our findings.

```
Sleep quality           1.000000
Time in bed             0.647258
Start time in second    0.192031
Start in hour           0.192031
End in hour             0.163342
End time in second      0.163342
Activity (steps)       -0.136605
Wake up                      NaN
Heart rate                   NaN
Name: Sleep quality, dtype: float64
```

Fig.5. Correlation between features to sleep quality numbered

Figure 5 provided above allowed for a detailed examination of correlations within the dataset. It revealed that sleep quality has the highest correlation with the amount of time spent in bed, while showing little to no correlation with other features. Based on this finding, we determined that these irrelevant features should be removed from the data as they do not contribute significantly to measuring sleep quality. Additionally, the "heart rate" feature in the dataset raised uncertainties as its values' interpretation was unclear, and it had the highest number of missing values. And so, we exclude the "heart rate" feature from the dataset, ensuring the integrity of the analysis.

By removing irrelevant features and excluding the uncertain "heart rate" variable, the dataset has been refined for a more focused analysis of sleep quality. The correlations presented in the figure provide valuable insights, emphasizing the strong association between sleep quality and time spent in bed. This streamlined dataset will facilitate reliable and meaningful results, enabling a deeper understanding of the factors influencing sleep quality.

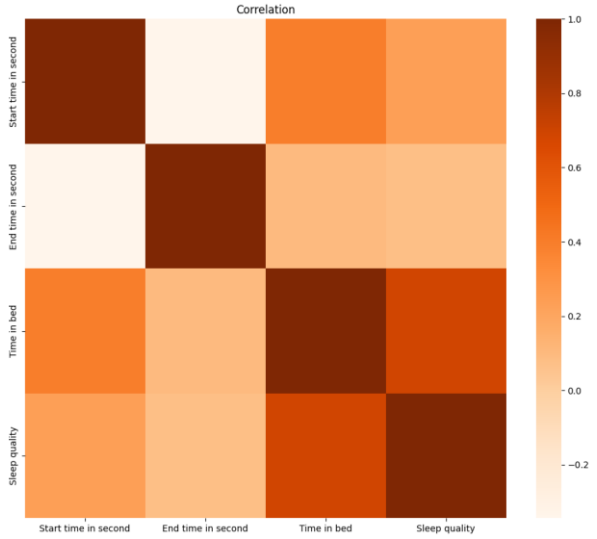

Fig 6. Correlation heatmap between features after extraction

After identifying the key features that correlate with sleep quality, we generated a new heatmap to visualize the relationships. From Figure 6 and accompanying table in Figure 5, it is clear that the most influential factor on sleep quality is "time in bed". And so, our calculations and algorithm development will primarily revolve around this metric, as it holds the highest impact in accurately assessing and improving sleep quality.

2.3 Machine Learning models

Six different machine learning models were trained and evaluated for their performance in predicting sleep quality. The models included a support vector machine (SVM), Random Forest (RF), Multi Layer Perceptron (MLP), Linear Regression, Gradient Boosting, and KNN. The models were trained on a subset of the data and evaluated on a separate validation set using 10-fold cross-validation. Several machine learning models were used in our research : Random Forest Regressor, Linear Regression, and Neural Network Regression. These machine learning models were trained and tested using different hyperparameters that suit the models best. In our code, we train the models using X as "time in bed" value and Y for our test prediction as the "sleep quality" value.

2.4 Model Evaluation

The performance of each model was evaluated based on several metrics, R-squared score, Mean squared error, and Mean absolute error. Both training and testing metrics of each model are evaluated to gauge out each model performance.

2.5 Further Analysis

Comparison is done between models to find which model achieves a higher compatibility score in this research. We are looking for a model that is suitable for the data that we gathered, not underfitting nor overfitting.

III. Results and Discussions

We present the results of the performance comparison among the different machine learning models. The R2 scores obtained for each model are reported, indicating their respective goodness of fit. We analyze and interpret the findings from the model performance comparison. We identify the models that achieved higher R2 scores, indicating their superior fit to the sleep data.

Machine Learning model	Training R-squared score	Test R-squared score
Linear Regression	0.46259	0.44688
MLP Regressor	0.48108	0.45711
KNN Regressor	0.66459	0.34049
Gradient Boosting	0.59041	0.39859
SVM	0.27630	0.25607
Random Forest	0.88541	0.32325

Fig 7. Table of Comparison between models compatibility

Based on the experimental results and analysis, we provide recommendations for selecting an appropriate machine learning model for sleep data analysis. We highlight the models that demonstrated superior performance and discuss their potential applications in sleep research and interventions. Based on figure 7, MLP Regressor has similar value of training and testing R-squared values. Meaning the model has a better generalization performance than other models. The training score is not significantly higher than the test meaning that the model is not overfitting to the data provided as stated in a study by [16].

Meanwhile, when we look at Random Forest, the model has a high value of training score of 0.88541 while the test score is significantly lower at 0.32325, meaning the model is overfitting though the model can understand the pattern better than others but cannot make generalizations on new unseen data.

Therefore, the MLP Regressor is recommended as a suitable machine learning model for sleep data analysis due to its balanced performance and generalization capabilities. Though it is important to note that these recommendations are based on the specific dataset and experiments conducted in our study. However, we can further evaluate the performance of the MLP model with more various datasets to verify its robustness. Overall, the MLP Regressor model is more capable of understanding the pattern between the data and can make generalization better than other models.

IV. Conclusion

In conclusion, this study aimed to evaluate and compare the performance of various machine learning models for sleep data analysis. The models were assessed based on their training and test R-squared scores, which provided insights into their goodness of fit and generalization capabilities. Among the models examined, the Random Forest model exhibited the highest training R-squared score, indicating a strong fit to the training data. However, its lower test R-squared score suggested potential overfitting, highlighting the need for caution when interpreting its performance on unseen data. So, we recommend the MLP Regressor model for this specific dataset in our study for its good generalization performance.

There are several limitations in our study. Lack of features included in the test make the models have a harder time to recognize patterns in the datasets. Datasets that we processed were also very limited, with less than 1000 datasets. In the future, we need to consider using more features extracted from data to make our results better. We hope that this research contributes to the field of sleep research and interventions by guiding researchers, clinicians, and sleep experts in selecting suitable machine learning models for sleep data analysis.

V. References

[1] Garbarino, S., Lanteri, P., Bragazzi, N.L. *et al.* Role of sleep deprivation in immune-related disease risk and outcomes. *Commun Biol* **4**, 1304 (2021). https://doi.org/10.1038/s42003-021-02825-4

[2] Kılıç, O., Saylam, B., & İncel, Ö. D. (2023). Sleep Quality Prediction from Wearables using Convolution Neural Networks and Ensemble Learning. *arXiv preprint arXiv:2303.06028*.

[3] Ravan, M. (2019). A machine learning approach using EEG signals to measure sleep quality. *AIMS Electronics and Electrical Engineering*, *3*(4), 347-358.

[4] Onyutha, C. (2020). From R-squared to coefficient of model accuracy for assessing" goodness-of-fits". *Geoscientific Model Development Discussions*, 1-25.

[5] D. H. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," Journal of Applied Science and Technology Trends, vol. 01, no. 04, pp. 140-147, 2020, doi: 10.38094/jastt1457.

[6] D. Radulović and D. Negovanović, "Gait Speed Prediction Based on Walking Parameters Using MLPRegressor", International Scientific Student Conference RI-STEM-2021 Rijeka, Croatia, pp. 21-26, 2021.

[7] Cai, L., Yu, Y., Zhang, S., Song, Y., Xiong, Z., & Zhou, T. (2020). A sample-rebalanced outlier-rejected $ k $-nearest neighbor regression model for short-term traffic flow forecasting. *IEEE access*, *8*, 22686-22696.

[8] Cai, J., Xu, K., Zhu, Y., Hu, F., & Li, L. (2020). Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. *Applied energy*, *262*, 114566.

[9] Dewi, C., & Chen, R. C. (2019). Random forest and support vector machine on features selection for regression analysis. *Int. J. Innov. Comput. Inf. Control*, *15*(6), 2027-2037.

[10] Segal, M. R. (2004). Machine Learning Benchmarks and Random Forest Regression. UCSF: Center for Bioinformatics and Molecular Biostatistics.

[11] Buettner R, Grimmeisen A, Gotschlich A (2020) High-performance diagnosis of sleep disorders: a novel, accurate and fast machine learning approach using electroencephalographic data. In: HICSS-53 proceedings, pp 3246–3255.

[12] Park, K., Lee, S., Cho, S., Wang, S., Kim, S., & Lee, E. (2019). Sleep prediction algorithm based on machine learning technology. *Eur Neuropsychopharmacol*, *29*, S514.

[13] Mahesh, "Machine Learning Algorithms - A Review," International Journal of Science and Research (IJSR), vol. 9, no. 1, Jan. 2020.

[14] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623.

[15] Danagerous, D.D (2022). Sleep Data. Kaggle. Retrieved from https://www.kaggle.com/datasets/danagerous/sleep-data

[16] Koehrsen, W. (2018). Overfitting vs. underfitting: A complete example. *Towards Data Science*, pp. 1-12.