Journal Geuthee of Engineering and Energy (JOGE)
Vol. 5, No. 1, May 2026, pp. 54 - 61
ISSN: 2964-2655, DOI: 10.52626/joge.v%vi%i.83

Predicting sleep quality using random forest on sleep health and lifestyle data

Naia Az-Zahra[1], Deri Latika Herda[2]
[1,2]Telecommunication Engineering, Polytechnic State of Padang, Indonesia

**Corresponding Author: naiaazzahra1906@gmail.com*

ABSTRACT

Sleep is an important physiological process that plays a role in maintaining the balance of biological and psychological functions. Lifestyle changes, such as high stress levels and a lack of physical activity, can affect a person's sleep quality. This study aims to analyze the influence of health and lifestyle factors on sleep quality and to develop a predictive model for sleep quality using the Random Forest algorithm. This study uses the Sleep Health and Lifestyle dataset with a classification approach into two categories, namely Ideal Sleep and Non-Ideal Sleep, determined based on sleep duration parameters referring to the concept of a U-shaped relationship and the sleep duration recommendations from the National Sleep Foundation. The data were processed through preprocessing and class imbalance handling using the SMOTE method, then split into training and testing data. The Random Forest model was built through hyperparameter tuning and evaluated using accuracy and Area Under the Curve (AUC) metrics. The results show that the Random Forest model achieved good classification performance with an Accuracy of 91.26%, Precision of 91.78%, Recall of 91.26%, and F1-Score of 91.30%. In addition, the model obtained an Area Under the Curve (AUC) value of 0.962, indicating very good classification capability. Based on the Feature Importance analysis results, the features with the greatest influence on sleep quality are Heart Rate, Stress Level, Physical Activity, and Daily Steps. The findings indicate that the combination of the SMOTE method and Random Forest is effective for predicting sleep quality based on health and lifestyle factors.

Keywords: Sleep Quality, Machine Learning, Random Forest, Prediction

ABSTRAK

Tidur merupakan proses fisiologis penting yang berperan dalam menjaga keseimbangan fungsi biologis, dan psikologis. Perubahan gaya hidup seperti tingginya tingkat stress, kurangnya aktivitas fisik dapat mempengaruhi kualitas tidur seseorang. Penelitian ini bertujuan untuk menganalisis pengaruh faktor kesehatan dan gaya hidup terhadap kualitas tidur serta membangun model prediksi kualitas tidur menggunakan algoritma Random Forest. Penelitian ini menggunakan dataset *Sleep Health and Lifestyle* dengan pendekatan klasifikasi dalam dua kategori, yaitu *Ideal Sleep* dan *Non-Ideal Sleep,* yang ditentukan berdasarkan parameter durasi tidur mengacu pada konsep *U-shaped relationship* dan rekomendasi durasi tidur dari National Sleep Foundation. Data diproses melalui tahap preprocessing dan penanganan ketidakseimbangan kelas menggunakan metode SMOTE, kemudian dibagi menjadi data pelatihan dan pengujian. Model Random Forest dibangun melalui hyperparameter tuning dan dievaluasi menggunakan metrik akurasi serta *Area Under Curve (AUC).* Hasil penelitian menunjukkan bahwa model Random Forest memperoleh performa klasifikasi yang baik dengan nilai Accuracy sebesar 91,26%, Precision 91,78%, Recall 91,26%, dan F1-Score 91,30%. Selain itu, model memperoleh nilai Area Under Curve (AUC) sebesar 0,962 yang menunjukkan kemampuan klasifikasi yang sangat baik. Berdasarkan hasil analisis *Feature Importance,* fitur yang paling berpengaruh terhadap kualitas tidur adalah Heart Rate, Stress Level, Physical Activity, dan Daily Steps. Hasil penelitian menunjukkan bahwa kombinasi metode SMOTE dan Random Forest efektif digunakan untuk memprediksi kualitas tidur berdasarkan faktor kesehatan dan gaya hidup.

Kata kunci: Kualitas Tidur, Machine Learning, Random Forest, Prediksi

1. INTRODUCTION

Sleep is a physiological process that is very important for humans because it plays a role in maintaining the balance of biological, psychological, social, and cultural functions. Poor sleep quality will cause various negative impacts, such as physical and mental health disorders like depression and anxiety. The quality of a person's sleep, whether good or bad, is influenced by various lifestyle factors, such as exercise, stress levels, and daily habits. Understanding the complex relationship between these various factors can provide insights in efforts to improve sleep quality [1].

SResearch on "The Effect of Physical Activity on Sleep Quality and Sleep Disorder" shows that physical activity has a positive relationship with sleep quality, indicating that the more physically active a person is, the better their sleep quality. It also shows that physical activity can reduce the severity of insomnia and various other sleep disorders [2]. These findings indicate that sleep quality is closely related to health and lifestyl conditions.

The development of technology in the field of artificial intelligence encourages the utilization of machine learning in health data analysis, including sleep quality prediction. Machine learning is a method capable of identifying patterns from the characteristics of the given data. In its development process, machine learning requires a sufficient amount of training data to build a model, then its performance is tested using testing data to determine the model's performance and accuracy level.[3].

Several previous studies have applied the Random Forest algorithm to classify sleep disorders into the categories of None, Sleep Apnea, and Insomnia. One study achieved an accuracy rate of 89.69%, with the best performance in the None category and a recall value reaching 96.08%[4]. These results indicate that the Random Forest algorithm has effective capabilities in processing complex health and lifestyle data.

However, most previous studies have focused on classifying types of sleep disorders rather than analyzing sleep quality based on ideal and non-ideal sleep categories. Therefore, research on sleep quality classification using sleep duration as a defining parameter remains limited. This study uses the concept of a U-Shaped relationship between sleep duration and health. Based on previous research, both too short (< 7 hours) and too long (> 9 hours) sleep durations are associated with an increased risk of health disorders[5]. Meanwhile, according to the recommendations of the National Sleep Foundation, the recommended sleep duration for teenagers and adults is in the range of 7-9 hours per night[6]. Therefore, this study aims to analyze the influence of health and lifestyle factors on sleep quality and to develop a sleep quality prediction model using the Random Forest algorithm. In addition, this study also evaluates the model's performance in classifying sleep quality into Ideal Sleep and Non-Ideal Sleep categories.

2. RESEARCH METHOD

This research focuses on data analysis related to sleep patterns and lifestyle, with the aim of understanding the factors that affect sleep quality and building a prediction model using the Random Forest algorithm to classify sleep quality into Ideal Sleep and Non-Ideal Sleep categories. The research stages are systematically designed to produce a valid prediction model. The series of research stages can be seen in Figure 1.

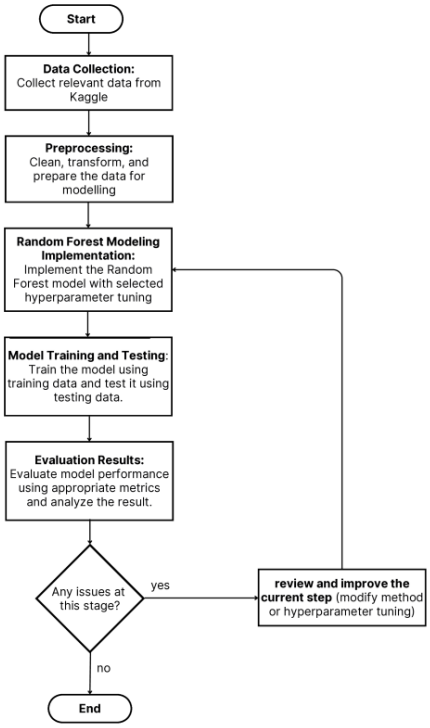

Figure 1. Research Stages

2.1. Dataset Description

The dataset used in this study is the Sleep Health and Lifestyle Dataset, which is publicly available on Kaggle. This dataset contains information about various lifestyle and health factors relevant to sleep quality [7]. The dataset has 514 rows and 13 columns. Missing values were detected in the 'Sleep Disorder' column.

The paramters used in this study include health and lifestyle factors, namely gender, age, occupation, BMI category, physical activity level, stress level, heart rate, daily steps, physical activity level, stress levels, BMI category, blood pressure, heart rate, daily steps, and the presence or absence of sleep disorders. Sleep quality was classified into Ideal Sleep and Non-Ideal Sleep categories based on sleep duration parameters.

2.2. Data Preprocessing

Before building the Random Forest model, a Preprocessing stage is required. This stage is very crucial to ensure the data is ready for machine learning modeling. The applied stages are useful for cleaning raw data, which often contains noise and missing values. Therefore, a series of steps are carried out as follows:

1. Removing the "Person ID" column
   Removing the 'Person ID' column using .drop('Person ID', axis=1). Person ID is a unique identifier that is not used for predictive analysis. If included in the model, it can cause overfitting.
2. Handling Zero or Missing Values
   Handling missing values in 'Sleep Disorder' by filling the missing values with 'No Disorder' so that there are no empty data that can interfere with the analysis process and model training.
3. Categorizing the Target Variable
   Categorizing sleep quality by considering the ideal sleep duration standard (7-9 hours per night)[6]. Then, categorical labels are converted into numerical representations using labelEncoder.
4. Implementation of the SMOTE method

SMOTE (Synthetic Minority Over-sampling Technique) is an oversampling technique used to address data imbalance. Unaddressed class imbalance can cause the model to have a high level of bias towards the majority class [8].

Before the application of SMOTE, the class distribution in the dataset was in an imbalanced condition. After the oversampling process using SMOTE, the number of data in each class became balanced, with each consisting of 234 data.

Table 1. Implementation of the SMOTE method

Class	Number of Data Before	Number of Data After
Ideal Sleep	234	234
Non-Ideal Sleep	177	234

5. Division of Training Data and Testing Data

The data is divided into 80% training data and 20% testing data using the train_test_split method. This division process aims to allow the model to be trained using one subset of data and tested on another subset that has not been used before, so that the evaluation of the model's performance on new data becomes more realistic and objective.

2.3. Modeling

Random Forest is one of the ensemble learning methods that consists of a collection of decision trees to perform the process of classifying data into a certain class. This algorithm uses the basic concept of decision trees, where input data is processed starting from the root up to the leaves, which are used to determine the classification result [9].

Random Forest is built randomly by generating many decision trees using different data subsets for each tree. Each tree produces a different prediction, then all these prediction results are combined to obtain a single final result through averaging [10]. To obtain the best parameters that can produce optimal model performance, hyperparameter tuning is performed. The determination of the best parameters is done by testing various combinations of parameters [10].

In this study, the hyperparameter tuning process was carried out using the GridSearchCV method. Before the tuning process was conducted, a baseline Random Forest model was first built using default parameters as a reference for the model's initial performance. The model was trained using data resulting from SMOTE oversampling (X_train_smote and y_train_smote).

The tuning process was performed by testing several parameters using the 5-Fold Cross Validation technique. The parameters used in the tuning process included n_estimators of 200, 300, and 500 to determine the number of decision trees in the forest, max_depth of 10, 20, and None to set the maximum depth of the decision trees, min_samples_split of 2 and 5 to determine the minimum number of samples before splitting a node, and min_samples_leaf of 1 and 2 to determine the minimum number of samples at each leaf. In addition, the max_features parameter used the sqrt and log2 methods to determine the maximum number of features at each node splitting process.

2.4. Evaluation Matrix

Performance testing is carried out using a confusion matrix. A confusion matrix is a table used to describe the performance of a classification method for which the actual values or classes are already known. It can help in evaluating the model's ability to correctly or incorrectly classify data in each class category [11]. The evaluation stage is conducted to determine how well the trained model predicts sleep quality:

1. Accuracy

Measuring how well the model predicts correctly:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2. Precision
   Shows how many positive predictions were correctly classified:
   $$Precision = \frac{TP}{TP+FP} \quad (2)$$
3. Recall
   Shows how many actual positive cases were successfully identified by the model:
   $$Recall = \frac{TP}{TP+FN} \quad (3)$$
4. F1-Score
   F1-Score is the harmonic mean between the values of precision and recall:
   $$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

True Positive (TP) is the Ideal Sleep class that can be accurately predicted by the model. False Positive (FP) is the Non-Ideal Sleep class but is predicted as Ideal Sleep. True Negative (TN) is the Non-Ideal Sleep class that is correctly predicted. False Negative (FN) is the Ideal Sleep class but is predicted as Non-Ideal Sleep.

3. RESULTS AND DISCUSSION

After the model is trained and tested on the test data, the model's performance is quantified through several evaluation metrics. This evaluation is carried out to determine the ability of the Random Forest algorithm in classifying sleep quality. In addition, feature importance analysis is also conducted to identify the features that have the most influence on the model's prediction results.

3.1. Model Performance

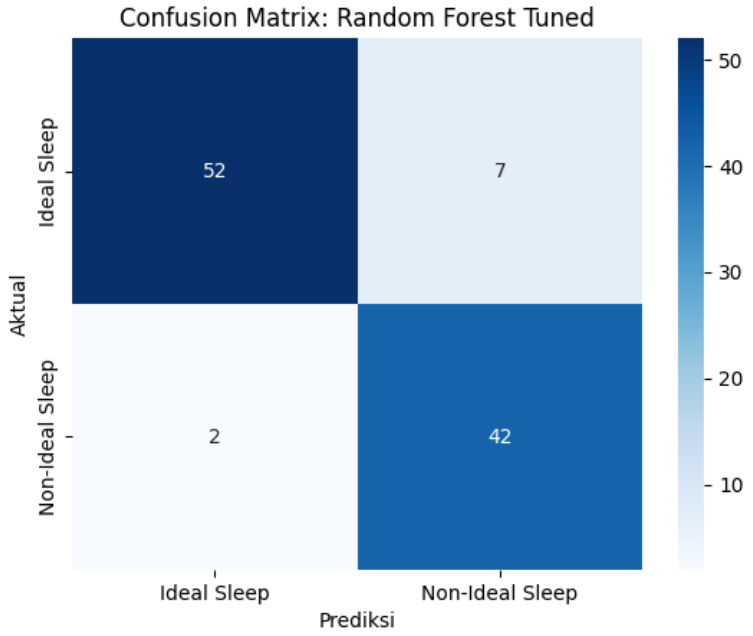

Figure 2. Confusion Matrix

In this study, *Ideal Sleep* was defined as the positive class, while *Non-Ideal Sleep* was treated as the negative class. Based on Figure 2. Confusion Matrix, the confusion matrix demonstrates that the tuned Random Forest model achieved strong classification performance. The model correctly predicted 52 Ideal Sleep instances (TP) and 42 Non-Ideal Sleep instances (TN). Meanwhile, 7 Ideal Sleep cases were misclassified as Non-Ideal Sleep (FN), and 2 Non-Ideal Sleep cases were incorrectly predicted as Ideal Sleep (FP). The dominance of correct predictions over misclassification cases supports the high evaluation results obtained, indicating that the proposed model effectively classified sleep quality categories.

Table 2. Test Results for Data Split 80%; 20%

Confusion Matrix	Random Forest (%)
Accuracy	91.26
Precision	91.78
Recall	91.26
F1-Score	91.30

Based on Table 2, the Random Forest model achieved an accuracy score of 91.26%. These results indicate that the model is able to classify sleep quality with good performance. The model was evaluated using test data that had previously been separated with a ratio of 80:20. The model performance evaluation results show an accuracy of 91.26%, precision of 91.78%, recall of 91.26%, and F1-Score of 91.30%.

3.1. ROC Curve

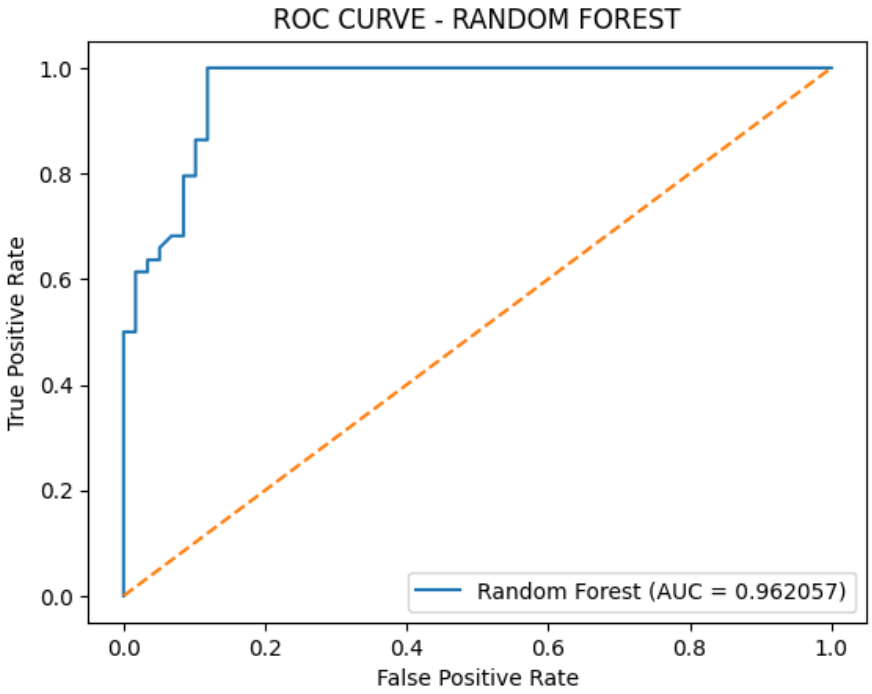

Figure 3. Receiver Operating Charactristic (ROC) Curve

Based on Figure 3. The results of the ROC Curve show that the Random Forest model demonstrates very good classification performance. The AUC value of 0.962 indicates that the model is able to distinguish between Ideal Sleep and Non-Ideal Sleep classes with a high level of accuracy. The closer the value is to 1, the more optimal the model's ability to perform the classification process.

3.3. Feature Importance Analysis

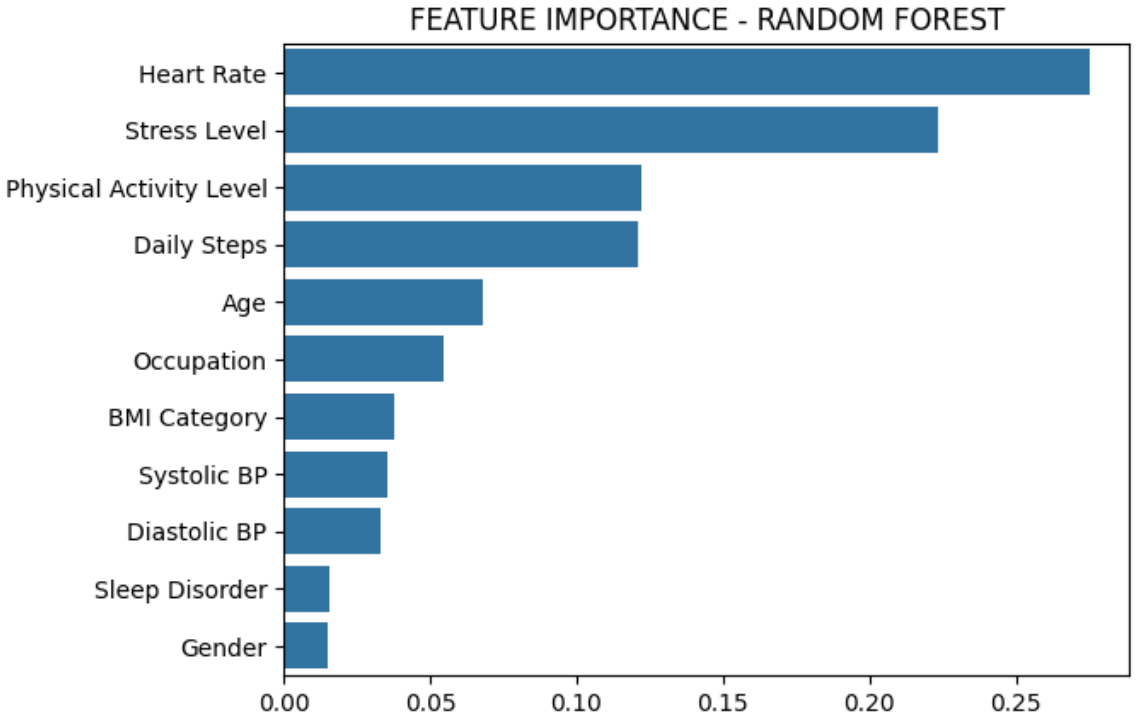

Figure 4. Feature Importance Analysis

Based on Figure 4. Results of the feature importance analysis, the feature that has the greatest influence on sleep quality prediction is Heart Rate, followed by Stress Level, Physical Activity Level, and Daily Steps. The high contribution of Heart Rate indicates that the physiological condition of the body has a strong relationship with a person's sleep quality. In addition, Stress Level also has a significant influence on the classification results. Physical activity factors, such as Physical Activity Level and Daily Steps, also make important contributions to the model. The better a person's physical activity, the better their sleep quality tends to be. Meanwhile, features such as Gender and Sleep Disorder have relatively small contributions compared to other features. These results show that the model considers physiological and lifestyle factors more than demographic factors in predicting sleep quality.

5. CONCLUSION

The results and discussion demonstrated that the proposed model was able to classify sleep quality into Ideal Sleep and Non-Ideal Sleep categories with high performance. The implementation of the Synthetic Minority Oversampling Technique (SMOTE) effectively addressed the data imbalance issue and contributed to improving the classification results. After hyperparameter tuning, the Random Forest model achieved an accuracy of 91.26%, precision of 91.78%, recall of 91.22%, F1-Score of 91.30%, and an AUC value of 0.962, indicating excellent predictive capability and classification performance. Based on the feature importance analysis, the factors that most influence sleep quality are Heart Rate, Stress Level, Physical Activity Level, and Daily Steps. These results reinforce that physiological factors and lifestyle have a relationship that significantly affects a person's sleep quality. Based on this, the findings of this study are in line with the research objective, which is to utilize machine learning to predict sleep quality based on health and lifestyle factors. The developed model has the potential to be integrated into health monitoring applications to automatically detect sleep quality. This study has several limitations.

The dataset was limited to the Sleep Health and Lifestyle dataset and sleep quality classification was mainly based on sleep duration parameters. In addition, this study only focused on the Random Forest algorithm without extensive comparison with other machine learning models, using a wider and more varied dataset, and adding additional health features so that the prediction model becomes more accurate and adaptive. The results of this study are expected to serve as a reference and source for future research development.

ACKNOWLEDGMENTS

The author expresses deepest gratitude to the supervising lecturer for the guidance, support, and invaluable input throughout the research process. The author also extends appreciation to Politeknik Negeri Padang for the academic support and facilities that have greatly contributed to the successful completion of this research. In addition, the author would like to thank the provider of the Sleep Health and Lifestyle Dataset which was used in this research.

REFERENCES

[1] M. Maulidah and N. Hidayati, "Prediksi Kesehatan Tidur Dan Gaya Hidup Menggunakan Machine Learning," *CONTEN Comput. Netw.*, vol. 4, no. 1, pp. 81–86, 2024, [Online]. Available:http://jurnal.bsi.ac.id/index.php/conten/article/view/4918%0Ahttp://jurnal.bsi.ac.id/index.php/conten/article/download/4918/1759

[2] M. A. Alnawwar, M. I. Alraddadi, R. A. Algethmi, G. A. Salem, M. A. Salem, and A. A. Alharbi, "The Effect of Physical Activity on Sleep Quality and Sleep Disorder: A Systematic Review," *Cureus*, vol. 15, no. 8, 2023, doi: 10.7759/cureus.43595.

[3] G. A. M. Ashfania, T. Prahasto, A. Widodo, and T. Warsokusumo, "Penggunaan Algoritma Random Forest untuk Klasifikasi berbasis Kinerja Efisiensi Energi pada Sistem Pembangkit Daya," Rotasi.

[4] N. Khasanah, D. U. Eka Saputri, F. Aziz, and T. Hidayat, "Studi Perbandingan Algoritma Random Forest dan K-Nearest Neighbors (KNN) dalam Klasifikasi Gangguan Tidur," *Comput. Sci.*, vol. 5, no. 1, pp. 17–25, 2025, doi: 10.31294/coscience.v5i1.5522.

[5] M. Du, M. Liu, and J. Liu, "U-shaped association between sleep duration and the risk of respiratory diseases mortality: a large prospective cohort study from UK Biobank," *J. Clin. Sleep Med.*, vol. 19, no. 11, pp. 1923–1932, 2023, doi: 10.5664/jcsm.10732.

[6] M. Hirshkowitz *et al.*, "National sleep foundation's sleep time duration recommendations: Methodology and results summary," *Sleep Heal.*, vol. 1, no. 1, pp. 40–43, 2015, doi: 10.1016/j.sleh.2014.12.010.

[7] L. Tharmalingam, "Sleep Health and Lifestyle Dataset." [Online]. Available: https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

[8] M. P. Pulungan, A. Purnomo, and A. Kurniasih, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 5, pp. 1033–1042, 2024, doi: 10.25126/jtiik.2024117989.

[9] A. Algiffary and T. Sutabri, "Analisis Random Forest Menggunakan Principal Component Analysis Pada Data Berdimensi Tinggi," *Indones. J. Comput. Sci.*, vol. 12, no. 2, pp. 284–301, 2023,[Online].Available:http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3135

[10] V. S. Prakash, S. N. Bushra, N. Subramanian, D. Indumathy, S. A. L. Mary, and R. Thiagarajan, "Random forest regression with hyper parameter tuning for medical insurance premium prediction," *Int. J. Health Sci. (Qassim).*, vol. 6, no. June, pp. 7093–7101, 2022, doi: 10.53730/ijhs.v6ns6.11762.

[11] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *Int. J. INFORMATICS Vis.*, vol. 7, no. March, pp. 258–264, 2023, [Online]. Available: www.joiv.org/index.php/joiv