COMPILER
Vol. 13, No. 1, May 2024, pp. 11~18
ISSN: 2252-2403 (Print), ISSN: 2549-2403 (Online)
Accredited Fourth Grade by Kemenristekdikti, Decree No:200/M/KPT/2020
DOI: 10.28989/compiler.v13i1.2184

Classification and Evaluation of Sleep Disorders Using Random Forest Algorithm in Health and Lifestyle Dataset

Wiwiek Widyastuty[1,*], Mochamamd Abdul Azis[2]
Department of Information Technology, Universitas Bina Sarana Informatika, Indonesia

Article Info

Article history:

Received April 11, 2024
Accepted May 8, 2024
Published May 31, 2024

Keywords:

Sleep Disorders
Insomnia
Machine Learning
Data Classification

ABSTRACT

Sleep is a basic human need, making up around one-third of our lives and being essential to the recovery of our physical well-being and general standard of living. However, poor sleep quality can interfere with these critical restorative processes, leading to disorders such as apnea and insomnia. These conditions not only impair daily performance but also have long-term health consequences. Furthermore, the challenges imposed by modern lifestyles have increased the prevalence of these sleep disorders, emphasizing the need for effective diagnostic tools. This research aims to harness the capabilities of Machine Learning (ML), specifically the Random Forest algorithm, to detect and analyze patterns indicative of sleep disorders in collected data sets. Random Forest is particularly suited for this task due to its ability to manage complex data sets by building multiple decision trees, thus creating a comprehensive and robust model for classifying sleep disorders. The findings of the study are promising, showing that the Random Forest algorithm can achieve a high level of accuracy in sleep disorder detection. The model demonstrated a test accuracy rate of 97.33%, with a precision of 96%, and a recall rate of 100%. Additionally, it achieved an F1-Score of 98% and a Kappa Score of 0.945, validating the reliability of this algorithm in producing precise classifications. This research offers significant insights into the patterns of sleep disorders and contributes to the development of targeted interventions aimed at improving sleep quality. Ultimately, this could significantly enhance the quality of life for individuals suffering from sleep disorders.

Corresponding Author:

Wiwiek Widyastuty,
Department of Information Technology
Universitas Bina Sarana Informatika
Jl. Kramat Raya No.98, RT.2/RW.9, Kwitang, Kec. Senen, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10450.
Email: *wiwiek.wwy@bsi.ac.id

1. INTRODUCTION

Everyone needs sleep; it's a natural state for the body and mind, and most people spend at least one-third of their lives asleep. Essential for physical health and a high quality of life, sleep facilitates numerous restorative functions in the human body, including memory consolidation, mental restoration, and the regulation of mood and behaviour, all of which are significantly influenced by the quality of one's sleep [1]. Sleep disorders, such as sleep apnea and insomnia, disrupt these essential functions. Sleep apnea disrupts a person's normal sleep cycle, affecting their daily routine and tasks. Insomnia prevents individuals from achieving the desired amount of sleep, significantly degrading both the quantity and quality of sleep [2]. The short-term effects of sleep disorders include a decreased ability to complete tasks on time.

This study is driven by the urgent need to address the challenges posed by sleep disorders in contemporary lifestyles, particularly affecting those who suffer from such conditions [3]. The prevalence of diseases related to sleep disorders is a significant concern, exacerbated by the impact of modern lifestyles and the widespread neglect of this fundamental need. The escalating risks associated with the increase in sleep disorders highlight

the importance of addressing this issue. One of the most important things for human survival is sleep, which is also very important for preserving general health and wellbeing [4].

The application of machine learning techniques in classifying sleep disorders is imperative not only to diagnose and treat these conditions effectively but also to enhance the overall quality of life and ensure the well-being of individuals in society [5]. Machine learning, a method often applied by many researchers, has introduced improved capabilities for automatic detection [6]. The data mining technique used for analysis in this study is the Random Forest algorithm, aimed at identifying patterns or characteristics indicative of sleep disorders. This algorithm classifies data using multiple decision trees, where each tree is created from a random vector of data. This approach, which involves random vectors in tree construction, helps in strengthening the model by increasing diversity between trees, thereby allowing more accurate identification of sleep disorders.

By utilizing machine learning, we can better understand sleep patterns, detect disorders with greater accuracy, and create focused interventions that can greatly enhance the quality of sleep and, in turn, the quality of life for those who are impacted by sleep disorders.

2. RESEARCH METHOD

This section focuses on the application of conventional machine learning methods, specifically the Random Forest algorithm, for classifying sleep disorders. The following part will describe the dataset used to evaluate the proposed algorithm, the performance metrics for model evaluation, and the feature importance techniques employed to compute the scores of the included features. The classification algorithm selected for this study is Random Forest, which will be briefly explained.

2.1. Random Forest

Bierman introduced the Random Forest algorithm in 2001. This algorithm is capable of handling two types of problems: classification and regression [7]. Random Forest is a development of the Classification and Regression Tree (CART) method, which uses the methods of bag or bootstrap aggregation and random feature selection. Bagging is a technique that can be applied to enhance the outcomes of an algorithm for classification. The ensemble approach is the foundation of the bagging method [8]. The Random Forest algorithm method can be divided into two stages; the first stage involves creating "k" trees to form a random forest, while the second stage uses the formed random forest to make predictions [9]. The process of applying the Random Forest method involves several steps, as described in [10] :

1. Initially, sample data is created by randomly selecting data points from the dataset, a process performed with replacement to allow the same data point to be selected more than once.
2. Subsequently, each sample is used to construct the i-th tree, where it represents the iteration ranging from 1 to k.
3. Steps 1 and 2 are repeated k times, corresponding to the desired number of trees to be built in the random forest.

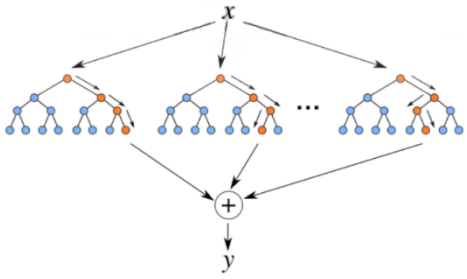

Figure 1. Random Forest

In the construction of decision trees using the CART (Classification and Regression Trees) method, the computation involves verifying information that explains how crucial each attribute is in classifying each node of the tree. Specifically, if we consider N as a node that separates the data classes D based on its attributes, this computation helps to measure how relevant or informative each attribute is in the process of data class separation. The node splitting process is performed by selecting the attribute that has the highest level of validation information. The formula used to calculate the level of validation information is as follows:

$$Gain\ (A) = Info\ (D) - Info(D) \quad (1)$$

To obtain the value of info(D), we can calculate it using formulas 2 and 3, which will produce the value of info A(D):

$$Info(D) = -\sum_{i=1}^{n} p_i log2(p_i) \quad (2)$$

Information:
n = number of target classes
pi = proportion of class i with respect to partition D

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \quad (3)$$

For attributes containing continuous or numerical values, it is necessary to determine the best splitting point to group values. The process of finding the best resolution begins with sorting the data. The median or mean of each pair of adjacent values is used as a potential splitting point. For example, if attribute A is a continuous attribute, then all values of A are sorted and the median becomes one of the potential splitting points. This can result in two or more partitions, where in this example v = 2 (with j=1 and 2) is the possible number of partitions[11].

2.2. 10-Fold Cross Validation

10-fold cross-validation is a technique that divides the data into two segments: a training set to train the model and a test set to evaluate it [12]. The method begins by dividing the dataset into training and testing data using 10-fold cross-validation, which involves allocating 80% of the data for training and 20% for testing. During the cross-validation process, data is split into n equal-sized partitions D1, D2, D3, ... Dn, and then the training and testing are performed n times. In the i-th iteration, partition D-i serves as the testing data, and the remainder forms the training data. Figure 1 illustrates the validation process with 10-fold cross-validation, where the dataset is divided into 10 parts—training and test data—and tests are performed 10 times [13].

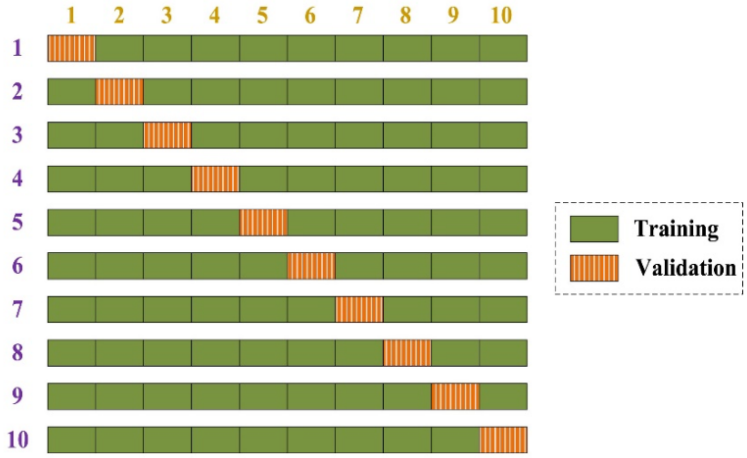

Figure 2. 10-Fold Cross Validation

2.3. Confusion Matrix

A confusion matrix is a method used to calculate the accuracy of data mining concepts. It contains an evaluation matrix that tests predictions to estimate the number of correct and incorrect classifications, giving measures of precision, accuracy and recall. Accuracy, also known as confidence, is the proportion of positive predictions that are true positives in real data. Recall or sensitivity is the proportion of true positive cases that are correctly predicted as positive [14]. This tool is essential for understanding the performance of a classification model, as it breaks down the results into true positives, false positives, true negatives, and false negatives, which are critical for gauging the model's ability to handle different classes accurately. It's particularly useful when the costs of different errors vary significantly.

Table 1. Model Confusion Matrix [15].

Correct Classification	**Clasified as**	
	+	-
+	True Positives (TP)	False Negatives
-	False Positives (FP)	True Negatives (TN)

The calculations of Accuracy, Precision and Recall based on the confusion matrix are as follows:

Accuracy Formula: $$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Precision Formula: $$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

Recall Formula: $\text{Recall} = \frac{TP}{TP+FN}$ (6)

3. RESULTS AND ANALYSIS

The presence of unresolved sleep disorders underscores the need for more detailed analysis. To address this, we have developed a classification model using the Random Forest Algorithm. This model is designed to analyze and detect sleep disorders more accurately. The workflow of our research in creating this model is illustrated in Figure 2 below, in the form of a flowchart.

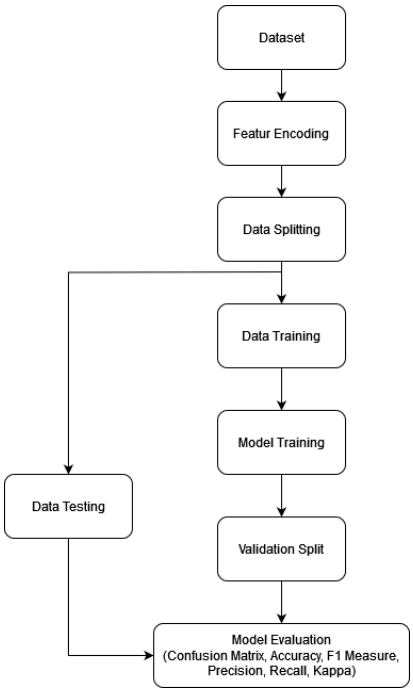

Figure 3. Research Stages

3.1. Dataset Analysis

The dataset used in this study is secondary data, publicly available and accessed from Kaggle. Consisting of 918 instances and 13 features, with a file size of 36 kB, this dataset contains electronic patient records, including basic physiological data and their medical history. The data has been stored in Google Drive after being downloaded from Kaggle.

The characteristics of this dataset include one binary attribute, which is Sleep Disorder, and four categorical attributes, namely Gender, Occupation, BMI Category, and Sleep Disorder. In addition, there are eight numerical attributes, which are Age, Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, Blood Pressure, Heart Rate, and Daily Steps. Thus, the attributes in this dataset consist of binary, numerical, and categorical types. Detailed information about the dataset attributes and their explanation is outlined in Table 2.

Table 2. Description of the Dataset Attribute

Attribute	Description
Person ID	A unique identifier for each individual
Gender	The gender of the person (Male/Female)
Age	The age of the person in years
Occupation	The occupation or profession of the person
Sleep Duration (hours)	The number of hours the person sleeps per day
Quality of Sleep (scale: 1-10)	A subjective rating of the sleep quality, ranging from 1 to 10
Physical Activity Level (minutes/day)	The number of minutes the person engages in physical activity daily
Stress Level (scale: 1-10)	A subjective rating of the stress level experienced by the person, ranging from 1 to 10
BMI Category	The BMI category of the person (e.g., Underweight, Normal, Overweight)
Blood Pressure (systolic/diastolic)	The blood pressure measurement of the person, indicated as systolic over diastolic pressure
Heart Rate (bpm)	The resting heart rate of the person in beats per minute

Daily Steps	The number of steps the person takes per day
Sleep Disorder	The presence or absence of a sleep disorder in the person (Insomnia, Sleep Apnea)

3.2. Data Cleaning and Preprocessing

Data cleaning and preprocessing are important steps in data processing before modeling.. This phase involves handling missing or inconsistent values in the dataset. This is important because missing or inconsistent data can lead to distortions in analysis and model outcomes. The process also includes converting data from one format to another that is more suitable for analysis. For example, converting text into numbers allows modelling algorithms to process data effectively. The goal is to produce a clean and structured dataset that is ready for use in statistical modelling or machine learning.

Table 3. Data Type After Transformation

Attribute	Data Type
Person ID	int64
Gender	object
Age	int64
Occupation	object
Sleep Duration	float64
Quality of Sleep	int64
Physical Activity Level	int64
Stress Level	int64
BMI Category	object
Blood Pressure	object
Heart Rate	int64
Daily Steps	int64
Sleep Disorder	object

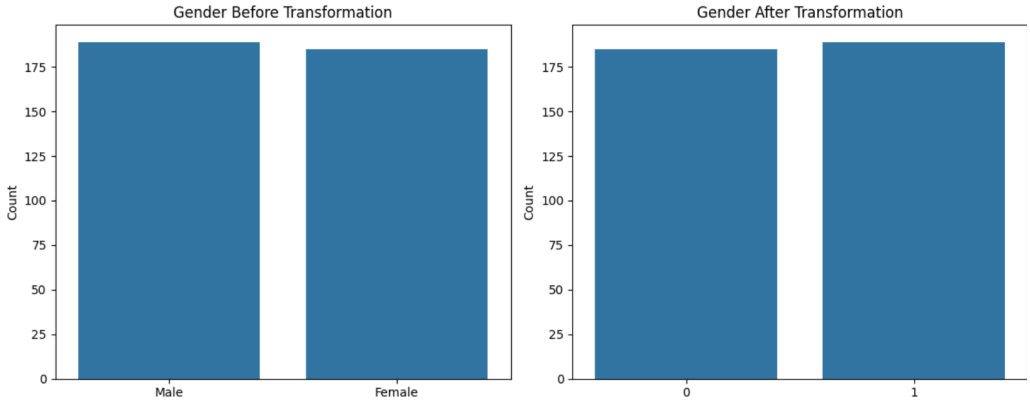

Figure 4. Transform categorical columns into numeric (*Encoder*)

3.3. Feature Engineering and Data Splitting

The next step is to share the information. The data is divided into two main sets: the training set and the test set. The training set is used to train the model and learn patterns in the data, while the test set is used to validate and test how well the model generalises to previously unseen data. This distinction is crucial to avoid overfitting, where a model performs well on training data but fails to accurately predict new data. The proportion of the split usually depends on the size and nature of the data, but an 80-20 (training-testing) split is most common. Additionally, cross-validation might be conducted, where the data is divided into several parts that are then used alternately as training and testing sets to ensure that the model is robust and consistent.

3.4. Model Training

Selecting a machine learning algorithm and training it with the training set is a crucial step. Here, hyperparameter tuning can also be performed to enhance the model's performance.

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

Figure 5. Train the model with the training set

3.5. 10-Fold Cross Validation

Before or after the training step, cross-validation is conducted to estimate the model's performance. This involves dividing the training data into several subgroups and repeatedly training the model to ensure that it is not overfitting and its performance is stable.

Table 4. Accuracy Score for Each Fold

Fold	Accuracy
1	0.7894736842105263
2	0.7368421052631579
3	0.9210526315789473
4	0.7631578947368421
5	0.9459459459459459
6	0.8108108108108109
7	0.918918918918919
8	0.6216216216216216
9	0.7027027027027027
10	0.918918918918919
Average	**0.8129445234708392**

3.6. Model Evaluation

Once the model is trained (and cross-validated), it is tested with a test set to assess accuracy and other evaluation metrics such as precision, recall, F1-score and confusion matrix.

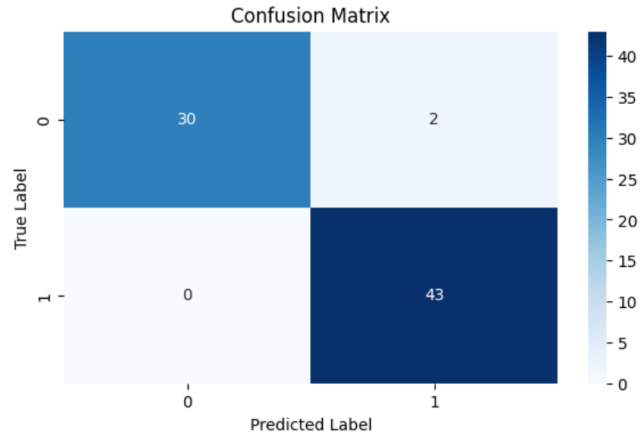

Figure 6. Evaluation Of The Confusion Matrix Model

Classification Labels:

1. Label 0: Insomnia – Represents observations that do not have Apnea but may suffer from Insomnia or other non-Apnea sleep conditions.
2. Label 1: Apnea – Represents observations diagnosed with Apnea, a medical condition characterized by periods where breathing stops or becomes very shallow during sleep.

Confusion Matrix Results:

1. True Negatives (TN): 30 observations – Correctly classified as not having Apnea (Insomnia).
2. False Positives (FP): 2 observations – Incorrectly classified as having Apnea when they do not.

3. False Negatives (FN): 0 observations – No instances where the model failed to identify actual cases of Apnea.
4. True Positives (TP): 43 observations – Correctly identified as having Apnea.

This structured analysis allows for a clear understanding of the model's performance in distinguishing between Insomnia and Apnea.

Table 5. Performance Independent Evaluation Table

Metric	Score
Accuracy	97.33%
Precision	96%
Recall	100%
F1-Score	98%
Kappa Score	0.945

Table 5 illustrates the performance of the Random Forest algorithm in model evaluation. The data presented shows that Random Forest delivered excellent performance on the test set. The algorithm achieved an accuracy rate of 97.33%, indicating that the model is highly effective in making correct predictions. A precision of 96% demonstrates that when the model predicts a positive label, it is almost always accurate. Perfect recall of 100% indicates that the model successfully identified all actual positive cases. The F1-Score, which is the harmonic mean of precision and recall, stands at 98%, showing a good balance between these two metrics. A Kappa Score of 0.945 indicates an almost perfect agreement between the model's predictions and the actual values, well above what could be expected by chance.

4. CONCLUSION

The study utilizing the Random Forest algorithm for classification demonstrates outstanding performance across various metrics on an independent test set. The model achieves a remarkably high accuracy rate of 97.33%, complemented by a precision of 96% and a recall of 100%. Such metrics not only highlight the model's efficacy but also its ability to consistently identify true positives without false negatives. The F1-Score of 98% further underscores an exceptional balance between precision and recall, suggesting that the model effectively harmonizes the trade-offs between these two critical metrics. Additionally, the Kappa Score of 0.945 is indicative of a significant agreement between the model's predictions and the actual outcomes, greatly surpassing typical expectations by random chance. This level of performance validates the robustness of the Random Forest model, confirming its potential as a reliable tool for practical applications in various fields that require precise classification capabilities. The results thus endorse the model's deployment in settings where high accuracy and reliability are crucial, pointing towards its applicability in more complex, real-world scenarios.

REFERENCES

[1] M. Zokaeinikoo, "Automatic sleep stages classification," 2016.

[2] Y. Maali and A. Al-Jumaily, "A novel partially connected cooperative parallel PSO-SVM algorithm: Study based on sleep apnea detection," in *2012 IEEE Congress on Evolutionary Computation*, 2012, pp. 1–8.

[3] Y. J. Kim, J. S. Jeon, S.-E. Cho, K. G. Kim, and S.-G. Kang, "Prediction models for obstructive sleep apnea in Korean adults using machine learning techniques," *Diagnostics*, vol. 11, no. 4, p. 612, 2021.

[4] A. Fauzi, R. Supriyadi, and N. Maulidah, "Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Forest," *J. Infortech*, vol. 2, no. 1, pp. 96–101, 2020.

[5] F. Thabtah, "Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment," in *Proceedings of the 1st International Conference on Medical and health Informatics 2017*, 2017, pp. 1–6.

[6] M. Mambang and A. Byna, "Analisis perbandingan algoritma c. 45, random forest dengan chaid decision tree untuk klasifikasi tingkat kecemasan ibu hamil," *Semnasteknomedia Online*, vol. 5, no. 1, pp. 1–2, 2017.

[7] L. Fadilah, "Klasifikasi Random Forest pada data imbalanced," Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta, 2018.

[8] N. K. Dewi, U. D. Syafitri, and S. Y. Mulyadi, "Penerapan Metode Random Forest Dalam Driver Analysis," in *Forum Statistika dan Komputasi*, 2011.

[9] M. L. Suliztia and others, "Penerapan Analisis Random Forest pada Prototype Sistem Prediksi Harga Kamera Bekas Menggunakan Flask," 2020.

[10] A. H. Primandari and others, "Implementasi Artificial Inteligence untuk Memprediksi Harga Penjualan Rumah Menggunakan Metode Random Forest dan Flask (Studi kasus: Rohini, India)," 2020.

[11] R. A. Haristu and P. H. P. Rosa, "Penerapan metode Random Forest untuk prediksi win ratio pemain player Unknown Battleground," *Media Inf. Anal. Dan Sist.*, no. 2, pp. 120–128, 2019.

[12] M. Adipa, A. T. Zy, and M. M. Effendi, "KLASIFIKASI EMAIL PHISHING MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR," *J. RESTIKOM Ris. Tek. Inform. dan Komput.*, vol. 5, no. 2, pp. 148–157, 2023.
[13] I. Nurjanah, J. Karaman, I. Widaningrum, D. Mustikasari, and S. Sucipto, "Penggunaan Algoritma Naive Bayes Untuk Menentukan Pemberian Kredit Pada Koperasi Desa," *Explorer (Hayward).*, vol. 3, no. 2, pp. 77–87, 2023.
[14] D. Apriliani, A. Susanto, M. F. Hidayattullah, and G. W. Sasmito, "Sentimen Analisis Pandangan Masyarakat Terhadap Vaksinasi Covid 19 Menggunakan K-Nearest Neighbors," *J. Inform. J. Pengemb. IT*, vol. 8, no. 1, pp. 34–37, 2023.
[15] D. Safitri, S. S. Hilabi, and F. Nurapriani, "Analisis Penggunaan Algoritma Klasifikasi Dalam Prediksi Kelulusan Menggunakan Orange Data Mining," *RABIT J. Teknol. dan Sist. Inf. Univrab*, vol. 8, no. 1, pp. 75–81, 2023.