Machine Learning Techniques for Sleep Disorder Classification and Sleep Quality Assessment

Darshan Laljani
Department of Computer Science
Pandit Deendayal Energy University
Gandhinagar, India
darshan.lce21@sot.pdpu.ac.in

Rahil Ganatra
Department of Computer Science
Pandit Deendayal Energy University
Gandhinagar, India
rahil.gce21@sot.pdpu.ac.in

Aniket Suthar
Department of Computer Science
Pandit Deendayal Energy University
Gandhinagar, India
aniket.sce21@sot.pdpu.ac.in

***Abstract* - This study represents a significant advance in healthcare by using machine learning techniques to predict sleep disorders such as insomnia and sleep apnea by analyzing data collected from wearable sensors. The primary aim is to use various factors – including gender, age, occupation, sleep duration, physical activity level, stress, body mass index (BMI), blood pressure, heart rate, and total steps per day – to develop a nuanced and personalized approach to quality improvement sleep. This study highlights the importance of lifestyle factors such as daily workload and exposure to technology, as well as environmental conditions, highlighting how they play a major role in the prevalence and severity of sleep disorders. By integrating these different data points, the study aims to clarify the complex interplay between these variables and sleep health. To achieve its goals, the study uses several sophisticated machine learning algorithms. These include but are not limited to the k-nearest neighbor (k-NN) classifier, logistic regression, counterfactual, chi2, and several other algorithms that are adept at handling large and complex data sets. The methodology involves careful data cleaning and transformation processes to ensure the integrity and usability of data collected from wearables and other sources.**

Index Terms – KNN, Logistic Regression , Sleep Disorder, Counterfactual Method

I. INTRODUCTION

Sleep plays a critical role in maintaining general health since it affects mental clarity, physical health, and quality of life. On the other hand, the global prevalence of sleep disorders, including insomnia and sleep apnea, is startlingly high and presents serious public health issues. Many people find traditional methods of diagnosing and treating sleep disorders to be inaccessible or inconvenient due to the frequently intricate and invasive processes involved. A growing number of people are interested in using wearable sensors and other forms of technology to monitor and evaluate sleep patterns in a more convenient and non-intrusive way in response to these difficulties.

With the development of wearable technology, there is now a rare chance to gather a wide range of physiological and activity data related to sleep health. In a real-world situation, variables like blood pressure, heart rate, amount of physical activity, and length of sleep may now be continuously tracked. With the use of cutting-edge machine learning techniques, our research seeks to utilize this data to create predictive models that not only provide highly accurate diagnoses for sleep problems but also provide guidance on possible therapies. In order to fully understand the combined impact on sleep quality, this study focuses on combining many data sets collected through wearable devices, including gender, age, occupation, stress levels, body mass index (BMI), and more. Through an in-depth evaluation of these variables, the research aims to identify patterns and links that point to sleep disorders, offering a complete method for forecasting illnesses such as sleep apnea and insomnia.

This study will also look into how environmental factors and lifestyle choices affect the health of sleep. The analysis will look at the relationships between sleep problems and everyday activities, workload, technology use, and environmental factors. Through the use of advanced techniques such as logistic regression and k-nearest neighbor (k-NN) classifiers, this study intends to establish a predictive framework that is proactive in tracking and improving sleep quality, in addition to being reactive.

II. LITERATURE REVIEW

The current study is centered on several bodily metrics, physical activity habits, and lifestyle behaviors. Previous research has shown that a number of variables, such as workload, stress, exposure to technology, and surroundings, have a major impact on the quality of sleep.[6] Our focus was on finding the perfect balance and selection of criteria to improve the accuracy of sleep duration and efficiency predictions.
The Pittsburgh Sleep Quality Index, insomnia, and sleep apnea are just a few of the sleep disorders that the research aims to predict.[5] The references include research addressing variables like stress, academic pressure, smartphone use, and other factors that affect university students' quality of sleep.[1]

This survey addresses the relationship between stress variables and sleep quality as well as the effects of sleep loss on cognitive function. Studies also explore the impact of social media use on adolescent self-esteem, anxiety, depression, and quality of sleep.[6] This varied research highlights the complex interplay

that shapes overall well-being and sleep quality and is influenced by lifestyle, environment, and health.[4]

Additionally, creating reliable prediction models may be hampered by the quantity and quality of data, especially when it comes to subjective factors like stress.

III. Problem Statement

The goal of this study is to create a machine learning model that can correctly categorize a person's sleep disorder into one of three groups: Insomnia, Sleep Apnea, or Healthy. A wide range of input parameters, such as occupation, age, gender, physical activity level, stress level, subjective sleep quality rating, length of sleep, BMI category, blood pressure, heart rate, and number of daily steps, will be utilized by the model.

The principal aim is to attain a minimum of 85% overall classification accuracy, while guaranteeing that the model produces comprehensible and interpreted predictions. This degree of precision is essential for delivering dependable and trustworthy information about people's sleep health, facilitating the early identification and prompt treatment of possible sleep disorders[2].

The predictability and interpretability of the model's outputs will enable a more thorough comprehension of the variables influencing sleep conditions, which will aid in the creation of individualized advice and focused treatment regimens[1].

The integration of data from various sources, feature engineering to find pertinent features and cut down on noise, managing class imbalances in the dataset, and maintaining privacy and regulatory compliance for sensitive health data are some of the major issues that need to be resolved. As sleep patterns and conditions can be dynamic and subject to a variety of external factors, addressing scalability and enabling real-time analysis for continuous monitoring and timely interventions is also essential [3].

The model will be incorporated into an approachable platform or application that is available to both individuals and healthcare professionals after it has been developed and validated successfully. This will enable all parties involved to take advantage of the model's potential for better analysis and management of sleep quality. The anticipated benefits include early detection of sleep disorders, personalized recommendations based on individual characteristics, and decision support for medical professionals in the diagnosis and treatment of sleep-related conditions [5].

In the present era, where stress, anxiety, and lifestyle factors significantly contribute to sleep disorders, the study is in line with the growing need for efficient sleep analysis tools. The prevalence of sleep-related problems has increased as a result of the fast-paced, blue light-emitting nature of modern life, widespread use of electronics, and disruption of natural sleep-wake cycles [6].

This study attempts to provide a comprehensive solution that addresses the complexities of sleep health by utilizing machine learning and data analytics. The study provides healthcare professionals and individuals with actionable insights through personalized recommendations, decision support, and accurate classification, ultimately leading to improved overall well-being and quality of life.

IV. Methodology

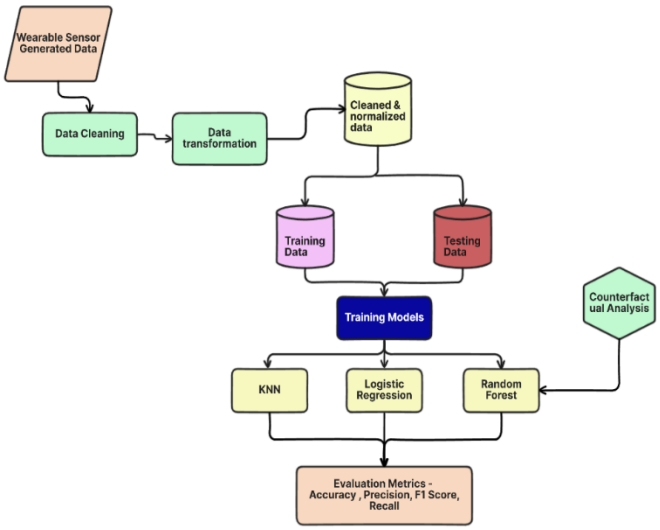

Figure 1. Flow diagram of methodology

In this study, we provide a machine learning-driven approach that uses data from wearable sensors to predict the two major sleep disorders namely Sleep pane and Insomnia based on physical and mental health factors.

A. Data Preprocessing

The study began with the deliberate collecting and creation of a comprehensive dataset from the wearable devices which included different types of body parameters like gender, age, occupation, sleep duration, physical activity level, stress level, BMI, blood pressure, heart rate, and total walking steps which are affecting the sleep quality of an individual, which are selected on the basis of Pittsburgh Sleep Quality Index parameters [3]. This initial stage plays an important role in building a solid foundation for the further study. This dataset preparation step was thoroughly carried out with the use of panda's package, various preprocessing and data cleaning tasks were performed like removing null values, eliminating duplicate values, finding correlations and handling various other anomalies in order to assure data quality and integrity, this number of cleaning procedures were carried out. The dataset's original examination used descriptive statistical methods to explain its main trends, dispersion, and general distribution, which helped to provide an initial grasp of the sleep quality data's properties [5]. In order to get insights and parameter composition of our sleep quality analysis data this step plays a crucial role in order to achieve the desired efficiency and output for the prediction to be made.

B. Model Selection and Training

For the predictive modelling task of sleep disorder prediction, three distinct algorithms were selected which are K-Nearest Neighbours (KNN), logistic regression, and Random Forest.

i. K- Nearest Neighbour

KNN classifier uses a voting process among the principle of 'learning by analogy' is emphasized while interacting with nearby things, a k-nearest neighbour classifier looks an n-dimensional vector space for the most comparable training information to the test information. Euclidean distance is used to evaluate closeness between test and preparing information sets.

$$D_{euc}(P, Q) = \sqrt{\sum_{i=1}^{n}(p - q)^2} \quad - (1)$$

Given 2 nodes P and Q in an n-dimensional vector space with P (p1, p2,… pn) and Q (q1, q2,..,qn), the distance between P and Q can be measured using (1)

ii. Logistic Regression

Logistic regression is a linear classification strategy that calculates the probability of a binary result using one or more predictor factors. Logistic regression was used in this study because it is straightforward, easy to interpret, and efficient when dealing with huge datasets. The logistic regression model was trained using the pre-processed dataset, and techniques like one hot encoding were utilized to avoid any prediction anomaly.

$$Y = e^{(\beta_0 + \beta_1 x)} / 1 + e^{(\beta_0 + \beta_1 x)} \quad - (2)$$

Using the input parameter x and biases β_0 , β_1 value of Y is calculated using (2)

iii. Random Forest

Random Forest is an ensemble learning method which generates a significant amount of decision trees during training and returns the class that is the basis for the classes (classification) or the average prediction (regression) of the individual trees. The Random Forest technique was chosen because it can handle data with high dimensions, capture complex relationships between features, and reduce overfitting. The model was trained using the pre-processed dataset, and hyperparameter adjustment was done to optimize its performance.

Each of these models were thoroughly trained on the pre-processed dataset, with hyperparameters optimized to enhance predictive performance. The models that were trained were then evaluated against standard metrics of performance to assess their usefulness in predicting sleep disorder results.

C. Predictive Analysis

To anticipate the occurrence of sleep disorders, with a focus on insomnia and sleep apnea. By utilizing machine learning methods such as K-Nearest Neighbours (KNN), Random Forest, and Logistic Regression, this study seeks to offer valuable insights for strategic planning and decision making in the prevention of sleep disorders. These algorithms were implemented using Python libraries such as scikit-learn, enabling rigorous training with historical data on sleep disorder occurrences [6]. A non-parametric classification technique called K-Nearest Neighbours (KNN) finds the majority class which among a data point's k closest neighbours in the feature space to predict the classification of that data point. In order to estimate the probability that participants in this study would suffer from sleep problems such insomnia and sleep apnea, logistic regression was also employed for further analysis [2]. We used the Random Forest model to forecast the occurrence of disorders of sleep based on characteristics like levels of stress, physical activity, and sleep duration. The model's performance was improved by tuning the hyper parameters like n_neighbors, metrics, weights, algorithm in case of KNN. All the three models were trained rigorously on the pre-processed data and with this three-model implementation, the comparison table (Table 1.) is prepared.

D. Counterfactual Methodology Integration:

The counterfactual approach produces "what-if" circumstances by changing the input features of a predictive model and assessing the resulting changes within the anticipated comes about. Within the context of sleep disorder forecast, the counterfactual strategy permits individuals to explore with way of life alterations or strategies that might lead to improved sleep health.

i. Methodology

We integrated the counterfactual technique into the Random Forest model by taking advantage of decision trees' inherent interpretability. For every data point in the dataset, we created counterfactual cases by altering the attribute values while leaving the outcome variable unchanged. These counterfactual instances illustrate different situations in which a person's lifestyles or activity undergoes change in an organized way.

ii. Implementation

After the model's predictions, the counterfactual incorporation was performed as a post-processing step within the Random Forest model. For those expected to have sleep disorders, we developed counterfactual instances by altering attributes related to lifestyle factors such as sleep duration, physical activity, and stress levels, body weight. Individuals may gain insights into the possible impact of many different lifestyle changes on their sleep health through assessing changes in predicted results for these counterfactual scenarios.

iii. Example Scenario

Consider an individual who is likely to have a sleep Problems based on their current lifestyle. The counterfactual method allows the individual to consider alternative scenarios such as increasing daily physical exercise or limiting the amount of screen time before bedtime. The associated changes in expected outcomes provide practical information for the individual to carry out targeted actions and improve their sleep quality.

E. *Data Output and Visualization*

The display of the data in visually appealing ways marked the methodology's conclusion. With great care, smart correlation matrix, plots, and heatmaps were created to understand the various input features affecting the dependent features. This allowed for the prediction of the sleep disorders in effective manner. Additionally, smart correlation matrix shows the visuals which summarizes the relationships between variables in our dataset. The final visuals were crucial in illustrating the strategic effects of the independent variables upon dependent variables and also some trends in the features incorporated for the prediction. Also, the output result of counterfactual methods helps the individual to enhance his/her sleep quality by altering the various input parameter like weight, stress levels, physical activity which we have included and can be controlled by the individual as well.

This study developed a methodological framework with a diverse, thorough and iterative strategy. Data collection and preparation, predictive analysis, counterfactual method integration was among the processes it covered. With the goal of promoting well-informed decision-making processes in predicting the sleep disorder based on parameters such as gender, age, occupation, sleep duration, physical activity level, stress level, BMI, blood pressure, heart rate, and total steps [6]. The integration of advanced predicting techniques and the multimodal approach underscores the study's commitment in order to predict the sleep disorders among individuals as well as providing them a way to enhance their sleeping conditions by altering various lifestyle changes given which are incorporated in the prediction using the counterfactual method.

V. VISUALIZATION AND KNN ANALYSIS

In the discipline of data science, the combination of KNN analysis with visualization is a powerful tool for extracting insights from large, intricate data sets. While KNN analysis offers an excellent method for solving regression and classification issues, visualization approaches using correlation matrix and plots offer a visual framework for comprehending data structures.

A. Smart Correlation matrix without masking

The correlation matrix (Figure 2.) illustrates the relationships between various parameters related to Quality of sleep and other physical and mental parameters. Each cell in the matrix contains a correlation coefficient, ranging from -1 to 1, indicating the strength and direction of the linear association between two variables.

Figure 2. Smart correlation matrix for various parameters for analysing relationship among the parameters.

We can determine which variables have a positive or negative correlation with one another by looking at the correlation matrix. Evaluating the relationships between many elements influencing sleep health and physical activity levels is made easier with the help of this information. For instance, there appears to be a strong positive association between "Sleep " and "Quality of Sleep," indicating that higher sleep duration may be linked to better sleep quality. On the other hand, an adverse correlation between "Heart Rate" and "Stress Level" suggests that, higher heart rates may be associated with higher stress levels.

B. Correlation matrix with masking

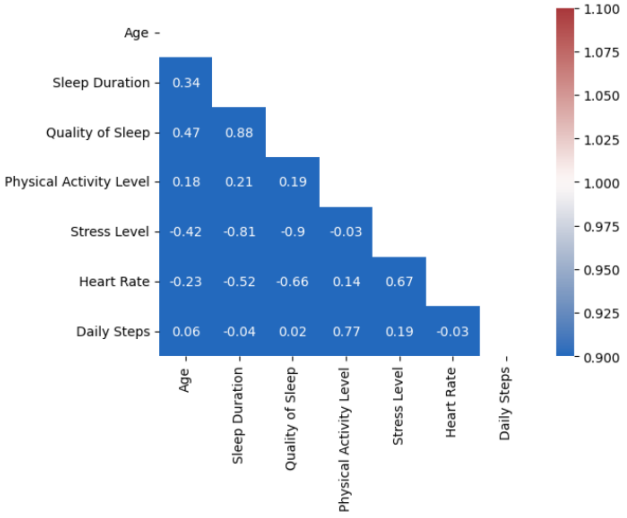

Figure 3. Correlation matrix without masking with the removal of redundant parameters.

The Correlation matrix (Figure 3.) incorporated matrix with masking to remove information in the upper triangle, and to display only the unique correlations in the lower triangle in order to provide clearer and more concise representation of the correlations.

C. *Average age of person with the disorder*

The graph (Figure 4.) shows the average age of people with three different sleep disorders: sleep apnea, insomnia, and none (presumably, no sleep disorder). The y-axis shows the average age and the x-axis shows the sleep disorder.
4 According to the graph, the average age of people with sleep apnea is around 50 years old. People with insomnia have an average age around 43 years old. The average person with no sleep disorder is around 39 years old.

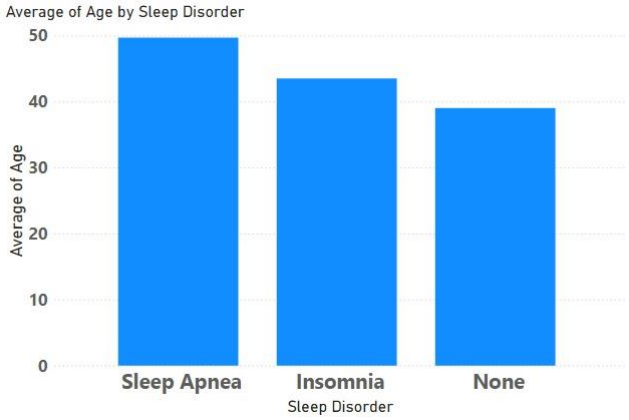

Figure 4. Plot displaying the average age of people with the sleep disorders to be predicted

D. *KNN Analysis table*

PARAMETERS						ACCURACY	Precision	Recall	F1-Score
n_neighbors	weights	algorithm	leaf_size	metric	n_jobs				
5	uniform	brute	30	minkowski	-1	0.88	0.90	0.88	0.88
5	distance	brute	30	minkowski	-1	0.89	0.89	0.89	0.89
5	distance	brute	30	euclidean	-1	0.89	0.89	0.89	0.89
5	distance	brute	30	manhattan	-1	0.89	0.89	0.89	0.89
5	distance	k-d_tree	30	manhattan	-1	0.89	0.89	0.89	0.89
5	distance	ball_tree	30	manhattan	-1	0.89	0.89	0.89	0.89
5	distance	auto	30	manhattan	-1	0.89	0.90	0.89	0.89
7	distance	auto	30	manhattan	-1	0.89	0.90	0.89	0.89
5	uniform	brute	30	minkowski	-1	0.88	0.90	0.88	0.88
9	distance	Kd_tree	30	manhattan	-1	0.89	0.89	0.89	0.89
9	uniform	ball_tree	50	manhattan	-1	0.89	0.89	0.89	0.89

Figure 5. Table with various parameter tuning for enhanced accuracy

An overview of the accuracy, precision, recall, and F1-score performance metrics of a K-Nearest Neighbours (KNN) classifier trained with various parameter configuration (Figure 5.) neighbours, weights, method, leaf_size, metric, and n_jobs are among the factors that were studied. The metrics that correspond to each row's representation of a particular parameter configuration are shown in the significant columns.

In a similar way, further settings for parameters with corresponding indicators of performance are provided in the rows that follow. This table provides a comprehensive guide to evaluate how various parameter values affect the KNN classifier's performance. It helps practitioners and researchers to find the best parameter combinations for classification jobs that result in the best predicted accuracy and efficiency.

VI. Results

The performance of the models implemented are shown in the following section. It offers the information of the model's evaluation metrics' data based on Accuracy, Precision, Recall and F1-score and the extent to which they match the study's objectives.

Table 1. Evaluation metrics (train data) analysis

Sr no.	ML algorithms	Accuracy	Precision	Recall	F1 Score
1.	KNN	0.93	0.94	0.93	0.92
2.	Logistic Regression	0.90	0.91	0.91	0.89
3.	Random Forest	0.89	0.90	0.90	0.90

Table 2. Evaluation metrics (test data) analysis

Sr no.	ML algorithms	Accuracy	Precision	Recall	F1 Score
1.	KNN	0.89	0.90	0.89	0.89
2.	Logistic Regression	0.92	0.92	0.91	0.91
3.	Random Forest	0.90	0.90	0.89	0.90

Using the result predicted by the logistic regression model the graph (Figure 6.) is prepared showing how much the number of actual classes and predicted classes are correctly predicted by the algorithm. From this plot we can analyze that the number of individual who are normal actually are less than that of predicted number which shows that somewhere some healthy persons are not being classified into correct sleep disorder. Also the same trend can be observed in case of Insomnia disorder prediction. This visual analysis is not merely a representation of data but also an analytical tool that provides a predictive overview of our model.

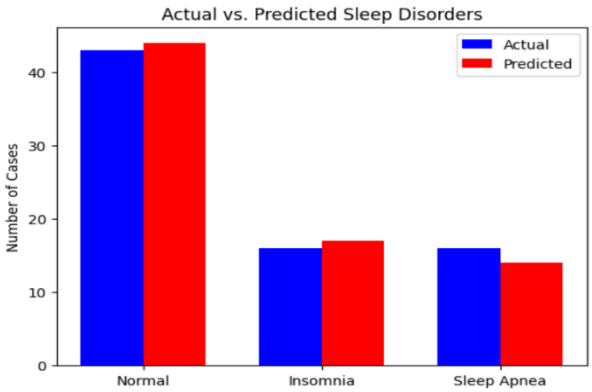

Figure 6. Histogram representing the number of predicted and actual classes by the Logistic Regressor model.

VII. Conclusion

In this study, we evaluated the effectiveness of three machine learning algorithms for sleep disorder prediction: Random Forest, Logistic Regression, and K-Nearest Neighbours (KNN). We managed to gain relevant results regarding the effectiveness of these models in classifying disorders of sleep based on a range of evaluation metrics by means of thorough experimentation and evaluation.

- Also, we can analyse that all three-model performed well in both training and testing phases and among them KNN model has the highest training accuracy while Logistic Regression model performed better in the testing phase. With the accuracy of 0.92 and F1 score of 0.91 the logistic regressor model was able to effectively predict the sleep disorders.

- The counterfactual method implemented over the prediction of Random Forest model identifies and recommends the parameters like Sleep duration, BMI category, Stress level as major affecting parameters on the Quality of sleep one can have. It is useful for providing.

In conclusion, our study's findings illustrate how well machine learning algorithms—KNN and Logistic Regression in particular—predict sleep disorders based on a range of characteristics. The counterfactual method's incorporation into Random Forest highlights the possibility of further advancements in predictive modelling approaches for the diagnosis and treatment of sleep disorders. These results open the door for more studies to improve the precision and usefulness of predictive models for the treatment of sleep disorders.

References

[1] L. Zhang et al., "Prediction of sleep quality among university students after analyzing lifestyles, sports habits, and mental health," Frontiers in Psychiatry, vol. 13, Aug. 2022, doi: https://doi.org/10.3389/fpsyt.2022.927619.

[2] W. Hidayat, Toufan Diansyah Tambunan, and Reza Budiawan, "Empowering Wearable Sensor Generated Data to Predict Changes in Individual's Sleep Quality," May 2018, doi: https://doi.org/10.1109/icoict.2018.8528750.

[3] Lo HM, Leung JHY, Chau GKY, Lam MHS, Lee KY, et al. (2017) Factors Affecting Sleep Quality among Adolescent Athletes. Sports Nutr Ther 2: 122. doi: 10.4172/2473-6449.1000122.

[4] K. Park et al., "Sleep prediction algorithm based on machine learning technology," European Neuropsychopharmacology, vol. 29, p. S514, 2019, doi: https://doi.org/10.1016/j.euroneuro.2018.11.763.

[5] Md. D. Manzar et al., "Validity of the Pittsburgh Sleep Quality Index in Indian University Students," Oman Medical Journal, vol. 30, no. 3, pp. 193–202, May 2015, doi: https://doi.org/10.5001/omj.2015.41.

[6] D. M. H. Kee et al., "Factors of Sleep Quality of University Students: A Comparison Between Malaysia and India," Asia Pacific Journal of Management and Education, vol. 4, no. 3, pp. 35–48, Nov. 2021, doi: https://doi.org/10.32535/apjme.v4i3.1264.