

Predicting Sleep Disorders Using Machine Learning: A Comparative Analysis

Sarah Goodyear and Abdallah Alsammani(✉)

Department of Mathematics, Jacksonville University, Jacksonville, FL 32211, USA
aalsamm@ju.edu

Abstract. Insomnia and sleep apnea represent sleep disorders that profoundly affect human health and well-being. This study examines how different machine learning (ML) approaches can determine sleep disorder risks based on physiological measurements alongside lifestyle and demographic data. The analysis involved a complete dataset that included sleep duration and quality measurements, physical activity levels, stress levels, cardiovascular health factors, and demographic information. The study conducted a thorough evaluation of multiple supervised classification algorithms, such as logistic Regression, decision tree, random forest, k-nearest neighbors (kNN), Gaussian naïve Bayes, multinomial naïve Bayes, Bernoulli naïve Bayes, support vector machines (SVM), and linear Regression. Decision tree, random forest, and kNN algorithms showed better predictive performance after preprocessing data thoroughly and selecting features through correlation analysis and cross-validation. The random forest model produced the best results with an accuracy rate of 94.38%, a precision level of 0.96%, a recall rate of 0.98% for the no sleep disorder class, and an AUC score of 0.98%. The feature importance analysis identified age, BMI category, blood pressure, and occupation as essential predictors for sleep disorders. The findings of this study demonstrate how machine learning methods enable the precise and timely detection of sleep disorder risks within clinical settings.

Keywords: Machine Learning · Sleep Disorders · Random Forest · Decision Tree · Logistic Regression · Health Informatics

1 Introduction

Sleep disorders constitute a significant public health challenge worldwide, affecting millions by lowering life quality and productivity while increasing healthcare expenses [6]. Two common sleep disorders that can result in serious health problems when untreated are insomnia and sleep apnea [2,17]. Insomnia involves challenges with falling asleep and staying asleep and reduced sleep quality, while sleep apnea causes repeated breathing interruptions during sleep [17].

New developments in technology alongside data science research have created fresh techniques for identifying and treating sleep disorders. Machine learning

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15648, pp. 106–121, 2025.
https://doi.org/10.1007/978-3-031-97000-9_7

(ML) enables systems to learn and self-improve from data through artificial intelligence techniques and has demonstrated significant potential in healthcare applications [7]. Machine learning methods provide potentially revolutionary new options for sleep disorder prediction and diagnosis beyond traditional diagnostic approaches [16].

While polysomnography (PSG) remains the primary diagnostic method for sleep disorders, it demands extensive resources alongside specialized laboratory facilities and trained technicians [12]. The diagnostic precision of PSG does not overcome its challenges, which include restricted accessibility, elevated costs, and scalability limitations [15]. The limitations of PSG can be overcome by combining machine learning techniques with accessible health metrics to create screening tools that identify at-risk individuals before needing extensive diagnostic tests [10].

Multiple research efforts have investigated different ML methods for classifying sleep disorders through support vector machines [11], neural networks [4], and decision tree models [19]. The latest findings demonstrate that machine learning algorithms successfully distinguish between various sleep disorders and typical sleep patterns by analyzing multiple physiological and behavioral datasets [3,9]. Despite recent advancements, performing comparative evaluations of ML algorithms is crucial to determine the best predictive methods for sleep disorders across various populations [14].

Our research focuses on creating several machine learning classification frameworks that predict sleep disorders by analyzing an extensive dataset containing health metrics like sleep duration and quality, physical activity, and cardiovascular indicators. Through the analysis of decision trees combined with random forests and k-nearest neighbors (kNN) algorithms, we determine which predictive models demonstrate the highest accuracy and which features impact classification performance [1,5,8].

The results of our study provide essential knowledge for medical practices and public health strategies. ML-driven tools that make detecting sleep disorders early and conducting screenings easier can enable faster interventions, which help decrease their harmful effects on personal health and healthcare systems [13]. Healthcare providers can develop personalized treatment plans and targeted prevention measures by identifying key predictive features.

2 Methods

2.1 Data Collection

The data used for this study is the Sleep Health and Lifestyle Dataset [18]. This dataset consists of synthetic data generated using algorithms to simulate results, allowing for unconstrained public analysis, and is stored in CSV format. There are 374 samples with data for 13 features relating to sleep, lifestyle, and physiological characteristics.

2.2 Data Exploration

Examining the contents of this dataset reveals that all samples are unique and contain no missing data. However, the Python data frame interprets the "None" in the "sleep disorder" column, representing no sleep disorder as clarified in the metadata, as a NaN value. Replacing these incorrectly labeled NaN values in the "sleep disorder" column with "No disorder" resolves this misinterpretation.

Boxplots were used to visualize the distribution of individual numeric features. This and individual statistical summaries reveal any possible skew or abnormalities. Age (Distribution 1) is spread approximately a bell curve from the minimum age 27 to the maximum age 59. Sleep duration (Distribution 2) is appropriately spread from 6.4 h to 8.5 h. Quality of sleep (Distribution 3) is left-skewed, ranging from a minimum of 4 to a maximum of 9, with the lower and upper quartiles falling between 6 and 8 and averaging 7. The physical activity level (Distribution 4) is evenly distributed, ranging from 30 to 90 min, with an average of 60 min. The stress level (Distribution 5) is relatively evenly spread from ratings 3 to 8, averaging 5, on a scale of 10. The heart rate (Distribution 6) is right-skewed, with a minimum of 65 to a maximum of 86, averaging 70 beats per minute. Daily steps (Distribution 7) are slightly left-skewed with a minimum of 3000, a maximum of 10000, a lower quartile of 5600, an upper quartile of 8000, and an average of 8000 steps daily. Person ID (Distribution 0) is an arbitrary label representing distinct subjects and is irrelevant to any numeric analysis. Of the other seven features, none of the distributions have any notable anomalies that could reasonably lead to deviated results.

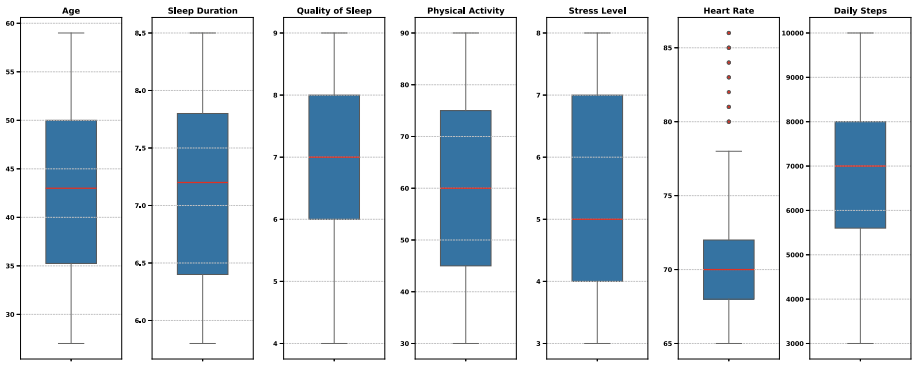

Fig. 1. Boxplots showing the distribution of seven health variables with median values (red lines), interquartile ranges (blue boxes), and outliers (red dots). The plots reveal typical values and variability for Age, Sleep Duration, Sleep Quality, Physical Activity, Stress Level, Heart Rate, and Daily Steps across the dataset. (Color figure online)

Similarly, bar plots were used to visualize the distribution of categorical features to examine possible anomalies. Gender (Distribution 8) is evenly distributed, with only four more male samples than female samples. Occupation

(Distribution 9) consists of nurses, doctors, engineers, lawyers, accountants, scientists, software engineers, sales representatives, and managers, in that order of prevalence. BMI category (Distribution 9) has the most occurrences of normal weight, followed by overweight, and then obese, making this most representative of a healthy-weighted population. However, normal weight is represented by both "Normal" and "Normal Weight," which can be combined to "Normal." Blood pressure (Distribution 11) has the most occurrences of 130/85, followed by 140/95, 125/80, and 120/80, accounting for most of these observations. This means the data is most representative of a population with slightly elevated blood pressure. Sleep disorder (Distribution 12) measures whether the subject has no disorder, sleep apnea, or insomnia, with no disorder being about three times as common in the population as either disorder. Combining "Sleep Apnea" and "Insomnia" to represent "Sleep Disorder" when a sleep disorder of any type is present allows for the prediction of the presence of a sleep disorder instead of focusing on specific types. Like the numeric features, no significant anomalies are present.

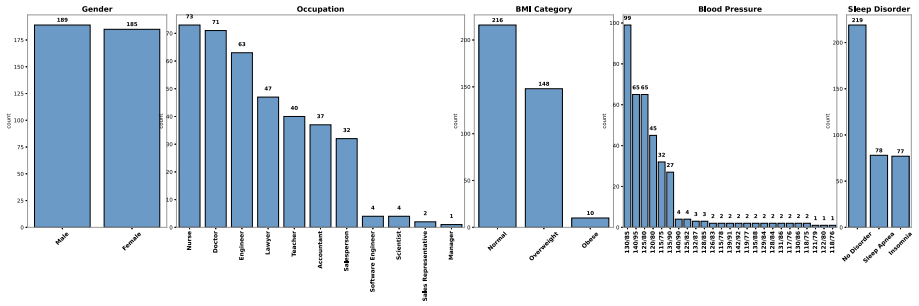

Fig. 2. The plots show count distributions Gendernder (balanced maleGendere representation), Occupation (dominated by healthcare and technical professions), BMI Category (mostly normal weight individuals), Blood Pressure (primarily clustering around 130/80 mmHg), and Sleep Disorder (majority with no disorder, with equal prevalence of sleep apnea and insomnia).

After cleaning the data, dummy variables were created for the categorical features so that numeric and categorical features could be normalized. They were normalized on a scale of 0 to 1. Normalization was more ideal than standardization for this study to limit the impact of the skew and any outliers present in the features by placing the values in a fixed range. Normalization also allows the data to be more easily interpreted by a wider audience and has the potential to perform better in a broader range of machine learning models. K-Nearest Neighbors (kNN) and many unsupervised techniques are more applicable when data is fixed on a bounded scale. Normalizing the data allows for a more thorough analysis of supervised techniques and prepares for a follow-up analysis and comparison with unsupervised techniques. With this data, several research goals

can be identified. First, cross-validation can be used to explore which supervised machine learning models have the most potential to predict the presence of a sleep disorder most accurately. The three models with the highest potential can be selected using these results to train more robust models. This can be used to gauge better what machine learning model can most effectively predict the presence of a sleep disorder with lifestyle and physiological data. Next, a correlation matrix was developed to aid in feature selection. Person ID is included in the matrix, but these metrics should be disregarded in the analysis since this feature is an arbitrary value to distinguish uniqueness and holds no statistical significance. The warmer the tone, the more positive the correlation; the cooler the tone, the more negative the correlation. Age, occupation, physical activity level, stress level, BMI category, blood pressure, and heart rate are observed to have positive correlations with sleep disorders. ConverseGendernder, sleep gender derivation, quality of sleep, and daily steps are observed to have negative correlations with sleep disorders.

3 Machine Learning

3.1 Feature Selection

Correlation metrics 3 were compared with the target variable, sleep disorder, using the correlation matrix to determine which features to include. Any corre-

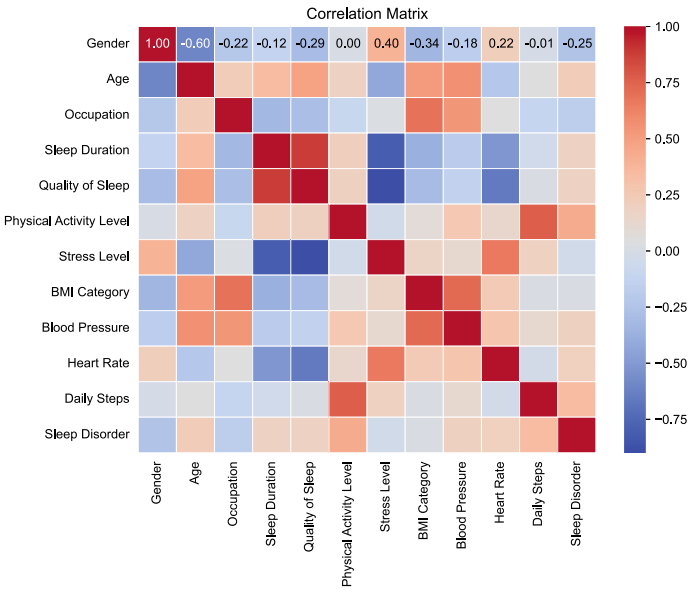

Fig. 3. The heatmap shows relationships between variables using color intensity (red for positive, blue for negative correlations). Notable relationships exist between sleep duration and quality, stress level and sleep metrics, BMI category, and blood pressure. (Color figure online)

lation with the person ID is disregarded. A threshold of 0.4 was selected, which accounts for the inclusion of four features: age, occupation, BMI category, and blood pressure.

3.2 5-Fold Cross Validation for Supervised Models

Nine supervised learning models were assessed using 5-fold cross-validation: decision tree, random forest, k-nearest-neighbors (kNN), Gaussian naïve bayes (GNB), multinomial naïve bayes (MNB), Bernoulli naïve bayes (BNB), logistic Regression, support vector machines (SVM), and linear regression model. Default parameters were set for each model before performing cross-validation. For the decision tree and random forest, the random state was set to 42. The number of neighbors for kNN was set to 5. The three naïve Bayes and the linear model types had no default feature set. Logistic Regression set the maximum iterations to 1000 and the random state to 42. SVM sets a linear kernel with an actual probability and a random state 42. For the cross-validation, a stratified k-fold was used, where there were five folds, shuffle was set to true, and random state was set to 42. A for loop was used to evaluate all nine models and store the accuracy, F1 score, and AUC-ROC, except the linear model, which only stored the correlation coefficient for later comparison. After evaluation, the models were ranked based on accuracy or correlation coefficient to determine the order of efficacy.

3.3 Models

The three models selected for training based on cross-validation results were the decision tree, random forest, and k-nearest neighbors (kNN). Parameter optimization was performed before training for each of these models. Each model was trained using the same four selected features and the same train-test split. The test size was 20% of the data, leaving 80% for training, and the random state was set to 42. After each model was trained, predictions were made using the test set, accuracy was computed, confusion matrices were produced, classification reports were printed, and auROC curves were created. These evaluation metrics were stored for each model for final comparison.

3.4 Logistic Regression

Logistic Regression is a statistical classification method that models binary outcome probabilities by examining the relationships between predictor variables and a binary response variable. The researchers used logistic Regression to differentiate between patients who have sleep disorders and those who do not based on specific physiological and demographic indicators.

The logistic regression model uses the logistic function to calculate the probability p that indicates whether a sleep disorder exists.

The probability p of having a sleep disorder is calculated using the logistic function as

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}} \tag{1}$$

p stands for the probability of the sleep disorder outcome with β_0 as the intercept and $\beta_1, \beta_2, \ldots, \beta_k$ as the coefficients for the predictor variables $X_1, X_2, \ldots, X_k$.

The logistic regression model transforms to a log-odds expression as follows:

The log-odds of the probability p equals the intercept β_0 added to the sum of products of regression coefficients β_i with predictor variables X_i for each i from 1 to k.

The left-hand side shows the logit transformation of probability p.

Demographic factors (age and sex) and physiological measurements (heart rate variability and sleep duration) were chosen as predictor variables for this study because previous sleep disorder research highlighted their significance. Model parameters ($\beta_0, \beta_1, \ldots, \beta_k$) were estimated through maximum likelihood estimation (MLE) using the training data subset after splitting the dataset into training and testing portions. The logistic regression model's performance was evaluated by applying the test dataset.

The classification accuracy and robustness of the logistic regression model were assessed through standard metrics, including accuracy, precision, recall (sensitivity), F1-score, and the receiver operating characteristic (ROC) curve with the area under the curve (AUC).

The logistic regression model showed excellent classification results by attaining high accuracy and maintaining balanced precision, recall, and F1-scores for both sleep disorder and no sleep disorder outcome classes. The detailed performance metrics will be discussed further in the Results section.

3.5 Decision Tree

The research team used a Decision Tree classifier to construct an interpretable model that classifies subjects according to their sleep disorder status. Decision Trees build hierarchical structures by recursively splitting data based on its feature values and producing decision nodes with leaf nodes representing classification results.

The dataset $\mathcal{D}$ comprises pairs (x_i, y_i) for $i = 1, \cdots, N$ where the feature vectors x_i belong to $\mathbb{R}^p$ and the class labels y_i are from the set $\{1, 2, \ldots, J\}$. The Decision Tree constructs distinct regions within the feature space $\mathbb{R}^p$ by implementing binary divisions at each node. The splits are chosen through impurity minimization to achieve the best separation between classes within each data subset.

The research applied the Classification and Regression Tree (CART) algorithm with the Gini impurity index for split evaluation. The mathematical definition of Gini impurity for a node dataset D is:

$$\text{Gini}(D) = 1 - \sum_{j=1}^{J} p_j^2, \tag{2}$$

Within node D, the variable p_j indicates the ratio of data points that fall into class j, while J indicates the number of unique classes. The splitting criterion selects partitions that reduce the aggregate weighted Gini impurities between child nodes.

$$\text{Weighted Gini} = \frac{|D_L|}{|D|}\text{Gini}(D_L) + \frac{|D_R|}{|D|}\text{Gini}(D_R), \tag{3}$$

D and D refer to the left and right child nodes after splitting, while $|D|$ depicts the total data points within the parent node.

The data was divided into training and testing subsets in a 70%/30% ratio, providing a stringent assessment of model generalization capacity. Grid search with 5-fold cross-validation was utilized on training data for hyperparameter optimization of maximum depth, minimum samples per split, and minimum samples per leaf to prevent overfitting.

The model performance evaluation on the test dataset involved multiple metrics, including confusion matrix analysis and precision, recall, F1-score calculations, and Receiver Operating Characteristic (ROC) curve examination. The precision and recall metrics were evaluated separately for every class. The F1-score is calculated as the harmonic mean of precision and recall.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

The ROC curve and the area under the Curve (AUC) were calculated to evaluate the model's classification performance across different classes.

The Decision Tree's interpretable structure enabled unambiguous identification of essential physiological and demographic predictors for accurate sleep disorder risk classification.

3.6 Random Forest

The ensemble learning Random Forest (RF) algorithm was the basis for classifying sleep disorders according to physiological and demographic data. The Random Forest method builds multiple decision trees during training before combining their results to improve accuracy while reducing overfitting.

The RF model consists of an ensemble that brings together multiple decision trees.

$$\{T_k(x, \Theta_k), k = 1, 2, \ldots, K\}, \tag{5}$$

The input feature vector is represented by x while Θ_k stands for independent, identically distributed random vectors that define each decision tree structure, and K denotes the total number of trees. The final classification output $\hat{y}$ results from:

$$\hat{y} = \arg\max_c \sum_{k=1}^{K} I(T_k(x, \Theta_k) = c), \tag{6}$$

The indicator function I(·) returns true when its argument matches the class label c.

The optimization of estimator number, maximum tree depth, and minimum samples per split and leaf was achieved through a grid search with 5-fold cross-validation during hyperparameter tuning. Predictive accuracy on the validation dataset determined the selection of optimal hyperparameters.

The machine learning pipeline started with feature selection, in which we identified key predictors from demographic data (age, sex, BMI), physiological parameters (heart rate variability, breathing rate, oxygen saturation), and lifestyle factors before proceeding to model training. The Gini impurity reduction criterion provided the mathematical definition to measure feature importance.

$$\text{Gini} = 1 - \sum_{i=1}^{J} p_i^2, \tag{7}$$

p_i indicates the fraction of samples from class i while J signifies the overall number of classes.

The model evaluation process used a hold-out test set to compute performance metrics, including confusion matrices, precision, recall, F1-score, and the AUC of the ROC curve. The interpretation of model performance became possible through confusion matrix analysis, which specified accurate favorable rates alongside true negatives and the amounts of false positives and negatives. The evaluation included precision and recall analysis for each class to achieve balanced prediction accuracy.

The RF model unified predictions from various decision trees through ensemble learning methods to produce reliable and stable classification outcomes. We implemented all analytical processes through Python's `scikit-learn` library.

4 Results

4.1 Cross Validation

A 5-fold cross-validation was performed to evaluate and compare the performance of nine supervised learning models. The models were ranked based on their classification accuracy, F1-score, and area under the receiver operating characteristic curve (ROC AUC), as shown in Table 1.

The top-performing models were the Random Forest and Decision Tree classifiers, both achieving an accuracy of 94.38%. The Random Forest model demonstrated superior discriminative performance with an ROC AUC of 98.00% and an F1-score of 91.5%, outperforming the Decision Tree, which had a slightly lower ROC AUC of 93.00% and an F1-score of 93.0%.

Logistic Regression also performed well, achieving an accuracy of 90.90%, an F1-score of 89.0%, and a high ROC AUC of 96.00%, indicating good overall classification ability despite its lower accuracy compared to the tree-based models.

Other models such as k-Nearest Neighbors (kNN), Support Vector Machines (SVM), and the various Naïve Bayes classifiers (Gaussian, Bernoulli, and Multinomial) showed moderate to strong performance. Still, they did not surpass the top three models. The kNN model achieved an accuracy of 93.30% and a ROC AUC of 92.22%, which were close but not competitive with the ensemble methods. Multinomial Naïve Bayes (MNB) had consistent scores around 86.90%, while Gaussian Naïve Bayes (GNB) and Bernoulli Naïve Bayes (BNB) hovered near 90.63% accuracy.

Linear Regression was the least effective model in this context. It did not output classification scores and was evaluated using the R^2 metric, which yielded a score of 64.36%, only marginally better than a random guess compared to classification-based models.

These results highlight the strength of ensemble learning and decision tree-based methods in classifying sleep disorders when trained on lifestyle and physiological data.

Table 1. Performance Comparison of Machine Learning Models for Sleep Disorder Classification and Regression

ML Model	Accuracy %	F1-score %	ROC AUC %	R2 score %
Random Forest	94.38	94.37	93.47	*NaN*
Decision Tree	94.38	94.37	93.91	*NaN*
kNN	93.30	93.30	92.22	*NaN*
Logistic Regression	90.90	90.91	93.38	*NaN*
SVM	90.63	90.65	92.70	*NaN*
BNB	90.63	90.65	90.34	*NaN*
GNB	90.37	93.80	93.36	*NaN*
MNB	86.90	86.91	86.03	*NaN*
Linear Regression	*NaN*	*NaN*	*NaN*	64.36

To evaluate the results of each model, a confusion matrix and area under the receiver operating characteristic curve (auROC) were developed. Classification reports showing precision, recall, and F1-scores were also created. These metrics give insight into how well each model can distinguish between classes and what classification types account for the error in predicting the test set.

4.2 Logistic Regression

The logistic regression classifier showed a strong ability to differentiate between patients with sleep disorders and those without. Figure 1 illustrates that the model attained high classification accuracy through various evaluation metrics. According to the confusion matrix in Fig. 1's left panel, the model demonstrated excellent predictive ability by accurately predicting 41 true negatives and 26 true

positives alongside eight incorrect classifications (2 false negatives and six false positives). The classification metrics analysis from Fig. 1's middle panel shows the model produced close precision results for both categories, with values of 0.87 for no sleep disorder and 0.93 for sleep disorder classes. The no-sleep disorder class achieved a higher recall rate of 0.95 against the sleep disorder class's 0.81, yet both classes showed balanced F1-scores of 0.91 and 0.87, respectively. The model's strong discriminative ability is confirmed by the receiver operating characteristic curve showcased in Fig. 1's right panel, which produces an area under the curve (AUC) value of 0.96, indicating exceptional classification performance. The evidence points to logistic Regression as a dependable tool for evaluating sleep disorder risks using specific physiological and demographic indicators.

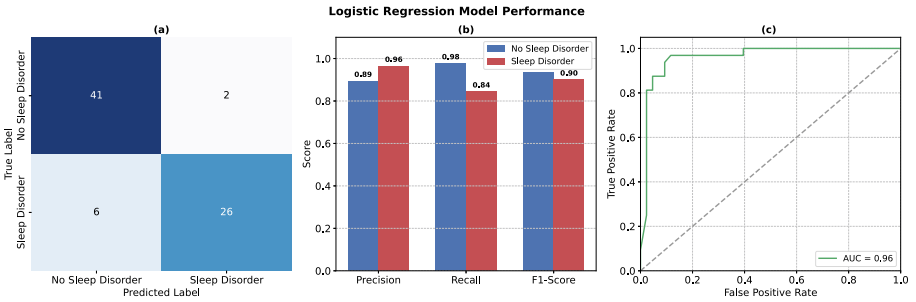

Fig. 4. Evaluation of Logistic Regression classifier. (a) Confusion Matrix; (b) Precision, Recall, F1-scores; (c) ROC Curve showing AUC = 0.96.

Figure 4 shows that logistic regression yields balanced performance across classes, with a particularly high AUC score of 0.96. The confusion matrix indicates a slight tendency toward false positives, while classification metrics suggest robust generalization capacity.

4.3 Decission Tree

The Decision Tree classifier demonstrated excellent predictive performance for sleep disorder classification, as illustrated in Fig. 2. he confusion matrix in the left panel displays how the model correctly recognized 41 subjects who do not have sleep disorders and 29 subjects with sleep disorders, while experiencing minimal errors with two false negatives and three false positives. According to the middle panel analysis, the precision metrics for both groups show excellent balanced classification performance between the no sleep disorder group at 0.93 and the sleep disorder group at 0.94. The model displayed excellent recall metrics for both class types, with scores of 0.95 for no sleep disorder and 0.91 for sleep disorder, which led to balanced F1-scores of 0.94 and 0.92, respectively. The ROC curve in the right panel demonstrates that the decision tree model has strong discriminative power with an AUC score of 0.93, significantly higher

than random chance performance. The optimized decision tree algorithm offers a dependable and understandable method for assessing sleep disorder risk with selected physiological and demographic features.

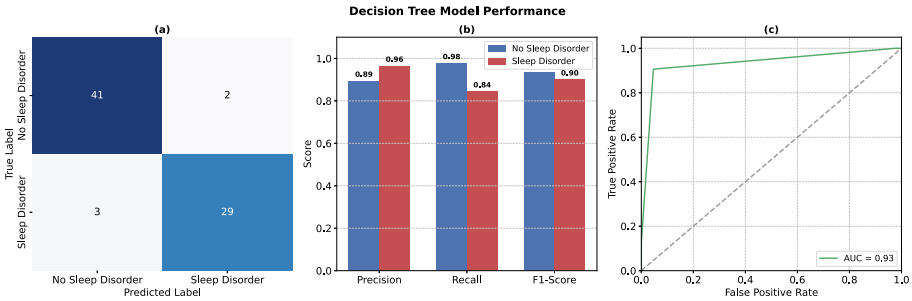

Fig. 5. Evaluation of Decision Tree classifier. (a) Confusion Matrix; (b) Class performance metrics; (c) ROC Curve with AUC = 0.93.

Figure 5 demonstrates the Decision Tree model's effectiveness in identifying sleep disorder status. he model shows excellent classification accuracy, with fewer misclassifications than logistic Regression. The AUC of 0.93 confirms reliable discriminative capacity.

4.4 Random Forest

The Random Forest algorithm demonstrated superior performance in sleep disorder classification. Figure 3 indicates that this ensemble learning approach delivered outstanding discriminative power across various evaluation metrics. The model has shown high accuracy through the confusion matrix in Fig. 3's left panel by correctly classifying 42 non-sleep disorder subjects and 27 sleep disorder subjects while maintaining minimal misclassification with one false negative and five false positives. The classification metrics analysis from Fig. 3, middle panel, demonstrates that different strengths complement each other across classes with a high precision of 0.96 in detecting sleep disorders and an excellent recall rate of 0.98 for the no sleep disorder class. The model achieved strong F1-scores of 0.93 and 0.90 for the two distinct classes. An AUC of 0.98 on the receiver operating characteristic curve (Fig. 3, right panel) demonstrates the model's outstanding capacity to separate classes, which approaches perfection in classification. The Random Forest ensemble method achieves dependable sleep disorder risk assessment by combining the predictive capabilities of numerous decision trees with selected physiological and demographic features.

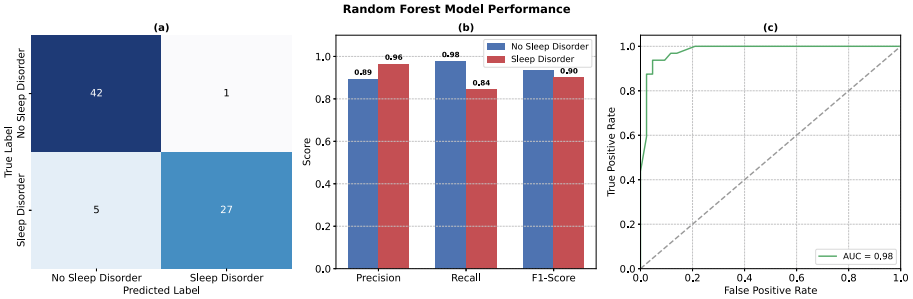

Fig. 6. Evaluation of Random Forest classifier. (a) Confusion Matrix; (b) Class-wise performance metrics; (c) ROC Curve with AUC = 0.98.

Figure 6 highlights the Random Forest model's superiority. It achieves the highest classification metrics across all models, with minimal false predictions. Its near-perfect AUC score of 0.98 confirms excellent reliability for clinical prediction use.

Table 2 presents the performance metrics of the three most effective models—Random Forest, Decision Tree, and Logistic Regression—after parameter tuning and final testing. All three models achieved high accuracy, F1-scores, and ROC AUC values, indicating strong predictive performance in classifying sleep disorder presence. The Random Forest model yielded the highest ROC AUC at 98.00%, demonstrating its superior ability to distinguish between classes. Although the Decision Tree and Logistic Regression models achieved the same accuracy of 94.38%, their F1-scores and ROC AUC values reveal nuanced differences: Decision Tree performed slightly better in F1-score (93.00%). At the same time, Logistic Regression showed stronger class separation with an ROC AUC of 96.00%. These findings confirm that ensemble methods and interpretable models can be effective, depending on the prioritized evaluation metric.

Table 2. Performance metrics of the top three optimized models

Model	Accuracy (%)	F1 Score (%)	ROC AUC (%)
Random Forest	94.38	91.5	98.0
Decision Tree	94.38	93.0	93.0
Logistic Regression	90.90	89.0	96.0

5 Discussion

This study examined the application of supervised machine learning algorithms to classify sleep disorders using the Sleep Health and Lifestyle Dataset. Through comprehensive data preprocessing, including cleaning, normalization, and feature selection via correlation analysis, we established a robust foundation for

model training and evaluation. Three of the nine models assessed were selected for optimization based on their strong cross-validation performance: Random Forest, Decision Tree, and Logistic Regression.

The Random Forest classifier achieved the most reliable and accurate results, attaining an accuracy of 94.38%, an F1 score of 91.5%, and an outstanding ROC AUC of 98.0%. This ensemble learning architecture, which integrates multiple decision trees, allowed the model to maintain generalization capability while minimizing overfitting. These findings confirm the Random Forest's ability to be a dependable predictive model for sleep disorder risk in clinical screening contexts.

The Decision Tree model also demonstrated high predictive accuracy (94.38%) and a superior F1 score (93.0%), with a slightly lower ROC AUC score of 93.0%. Despite its performance being marginally behind that of Random Forest, its interpretable structure makes it highly suitable for healthcare applications that require transparency and explainability. The ability to visualize and trace decision paths based on key features (e.g., age, BMI category, and blood pressure) enhances the model's utility for practitioners.

Logistic Regression offered strong discriminative performance with an ROC AUC of 96.0%, though it trailed overall accuracy (90.90%) and F1 score (89.0%) compared to the tree-based models. However, its simplicity, interpretability, and efficiency make it a practical option for initial screenings or in settings where computational resources are limited.

Overall, our results reinforce the value of supervised machine learning in the early detection of sleep disorders using accessible physiological and lifestyle indicators. The ensemble-based Random Forest model stands out for its robustness and classification power, while Decision Tree and Logistic Regression remain valuable options depending on clinical and operational needs.

6 Conclusion

The study confirms that supervised machine learning techniques can accurately classify sleep disorders when applied to demographic and physiological information. The Random Forest algorithm achieved superior predictive accuracy and demonstrated strong performance metrics, making it the preferred choice for clinical implementation. Decision Trees provide a transparent option for situations that require clear interpretation, but Logistic Regression continues to be an applicable choice because of its strong classification power.

Future research plans should extend to include real-world clinical data in the dataset, assess more physiological indicators, and implement unsupervised analysis methods for detailed patient subgroup studies. Amm is conducting ongoing clinical study validations to demonstrate these predictive models' practical applicability and widespread relevance. s essential

Acknowledgments. The authors would like to thank the Department of Mathematics at Jacksonville University for their generous support and funding assistance for the conference registration. Their encouragement and resources made this research and its dissemination possible.

Conflict of Interest. The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Mach. Learn. **6**(1), 37–66 (1991)
2. Alsammani, A., Gliske, S., Stacey, W.: Effect of sleep stage on high-frequency oscillations and artifacts. In: AES 2021 Annual Meeting Abstract Database. AESnet.org (2021), presented at the American Epilepsy Society Annual Meeting, November 22 (2021)
3. Alsammani, A., Stacey, W.C., Gliske, S.V.: Estimation of circular statistics in the presence of measurement bias. IEEE J. Biomed. Health Inform. **28**(2), 1089–1100 (2023)
4. Biswal, S., Sun, H., Goparaju, B., Westover, B.: An expert system for sleep apnea diagnosis using neural networks. J. Clin. Sleep Med. **14**(12), 2029–2040 (2018)
5. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
6. Chattu, V.K., Manzar, M.D., Kumary, S., Burman, D., Spence, D.W., Pandi-Perumal, S.R.: The global problem of insufficient sleep and its serious public health implications. Healthcare **7**(1), 1 (2018)
7. Deo, R.C.: Machine learning in medicine. Circulation **132**(20), 1920–1930 (2015)
8. Diaz-Uriarte, R., Alvarez de Andres, S.: Gene selection and classification of microarray data using random forest. BMC Bioinform. **7**, 3 (2006)
9. Goldstein, C.A., Berry, R.B., Kent, D.T., Kristo, D.A., Seixas, A.A.: Artificial intelligence in sleep medicine: background and implications for clinicians. J. Clin. Sleep Med. **16**(4), 609–618 (2020)
10. Goldstein, C.A., et al.: Artificial intelligence in sleep medicine: background and implications for clinicians. J. Clin. Sleep Med. **16**(4), 609–618 (2020)
11. Kang, J., Kim, J., Lee, H.G., et al.: Support vector machines for the diagnosis of insomnia. Sensors **17**(7), 1502 (2017)
12. Kapur, V.K., et al.: Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American academy of sleep medicine clinical practice guideline. J. Clin. Sleep Med. **13**(3), 479–504 (2017)
13. Khosla, A., et al.: Automated detection of sleep disorders using machine learning algorithms. J. Clin. Sleep Med. **14**(8), 1423–1430 (2018)
14. Liu, Y., et al.: Systematic analysis of long non-coding RNAs and mRNAs in the ovaries of Duroc pigs during different follicular stages using rna sequencing. Int. J. Mol. Sci. **19**(6), 1722 (2018)
15. Malhotra, A., Ayappa, I., Ayas, N., et al.: Diagnostic accuracy and limitations of polysomnography. Chest **153**(2), 333–340 (2018)
16. Mostafa, S.S., Mendonça, F., G. Ravelo-García, A., Morgado-Dias, F.: A systematic review of detecting sleep apnea using deep learning. Sensors **19**(22), 4934 (2019)
17. Sateia, M.J.: International classification of sleep disorders-third edition: highlights and modifications. Chest **146**(5), 1387–1394 (2014)

18. Tharmalingam, L.: Sleep health and lifestyle dataset. Kaggle. https://www.kaggle.com/datasets/uom190346a/sleep-health-andlifestyle-dataset. Accessed 8 Feb 2024 (2023)
19. Ting, H., Mai, Y.T., Hsu, H.C., Wu, H.C., Tseng, M.H.: Decision tree based diagnostic system for moderate to severe obstructive sleep apnea. J. Med. Syst. **38**, 1–10 (2014)