Sleep Disorder Prediction Using Machine Learning

Ganesh Ishwarchandra Ghonsikar[1], Ms. Rucha Ravindra Galgali[2]
[1,2]*Deogiri Institute of Engineering and Management Studies, Chhatrapati Sambhajinagar*

Abstract*—sleep disorder prediction involves identifying early signs that a person may develop problems with sleep, such as insomnia, sleep apnea, or restless leg syndrome. It relies on monitoring patterns in sleep duration, quality, and consistency. Factors like stress levels, irregular work schedules, lifestyle habits, and medical history can indicate higher risk. Physical symptoms such as excessive daytime sleepiness, difficulty falling asleep, or frequent awakenings are important clues. Regular tracking of these signals can help anticipate disorders before they become severe. Early prediction allows for timely lifestyle adjustments, medical consultations, and preventive measures to maintain healthy sleep. In this study, multiple machine learning algorithms are evaluated, including Random Forest, K-Nearest Neighbours, Support Vector Machine (RBF), Multilayer Perceptron (Neural Network), Decision Tree, XGBoost, Gradient Boosting, and Logistic Regression. The results show that tree-based ensemble methods, particularly Random Forest, Gradient Boosting, and XGBoost, delivered the best performance, achieving an accuracy of 94.7% and a weighted F1-score of 0.95, with balanced precision and recall across all sleep disorder categories. Among these, XGBoost emerged as the preferred model due to its strong generalization and robustness. Early prediction using such models can enable timely lifestyle adjustments, medical consultation, and preventive interventions to maintain healthy sleep.

Index Terms*—sleep disorder prediction, insomnia, sleep apnea, restless leg syndrome, sleep quality, early detection, preventive interventions, machine learning, Random Forest, XGBoost, Gradient Boosting, Decision Tree, K-Nearest Neighbours, Support Vector Machine (RBF), Multilayer Perceptron (Neural Network), Logistic Regression, sleep health.

I. INTRODUCTION

Sleep plays a vital role in maintaining physical health, mental stability, and overall well-being. However, sleep disorders often remain undiagnosed due to lack of awareness, high medical costs, and limited access to specialized clinical testing. Traditional diagnostic methods such as polysomnography are time-consuming, expensive, and unsuitable for large-scale or early screening.

With the increasing availability of health and lifestyle data, machine learning offers an effective approach for predicting sleep disorders using parameters such as sleep duration, stress level, physical activity, body mass index (BMI), blood pressure, occupation, and quality of sleep. These data-driven techniques can identify hidden patterns that are difficult to detect through conventional analysis.

This application was developed to provide an intelligent, affordable, and efficient system for early prediction of sleep disorders. By analyzing lifestyle and health-related attributes, the system aims to assist healthcare professionals and individuals in identifying potential sleep-related issues at an early stage. Multiple machine learning models are implemented and compared to determine the most accurate and reliable approach, demonstrating the practical use of machine learning in preventive healthcare and medical decision support.

II. OBJECTIVES AND NECESSITY

1. To explore how machine learning techniques can be used to accurately identify and predict various sleep disorders.
2.To compare existing models and find which methods work best.
3.To show why data-driven approaches are useful for fast, reliable, and non-invasive diagnosis.

III. LITERATURE REVIEW

A. RESEARCHERS CONTRIBUTIONS:

T. S. Alshammari highlighted that accurate sleep disorder classification is vital for improving health, as manual analysis is time-consuming and prone to errors. This study compares machine learning and deep learning models on the Sleep Health and

Lifestyle Dataset, using a genetic algorithm to optimize their parameters. Among k-nearest neighbours, SVM, decision tree, random forest, and ANN, the ANN achieved the highest accuracy of 92.92%, with strong precision, recall, and F1-score, outperforming other models.

Locharla Ravikumar et al., highlighted that sleep disorders like insomnia and sleep apnea significantly affect health and quality of life, making accurate diagnosis essential. Using the Sleep Health and Lifestyle Dataset, their Random Forest model with SMOTE-based class balancing achieved 96.70% accuracy in classifying Healthy, Insomnia, and Sleep Apnea cases. The study demonstrates the effectiveness of non-invasive, data-driven approaches as scalable alternatives to traditional diagnostic methods.

V. Shanmugapriya et al., explored the use of machine learning (ML) techniques to predict sleep disorders by analyzing physiological, behavioral, and demographic data. The study reviews models such as decision trees, SVMs, CNNs, LSTMs, and ensemble methods, using both public datasets and wearable device data. Key considerations include feature selection, data preprocessing, and hyperparameter optimization. Challenges like data privacy, model interpretability, and clinical integration are discussed. Findings indicate that hybrid ML approaches can enhance the accuracy and reliability of sleep disorder prediction.

B. ALGORITHMS USED IN PREVIOUS RESEARCH

1. Random Forest: Random Forest (RF) is an ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting. It uses bootstrapping to train each tree on a different subset of data, enhancing robustness. Random feature selection ensures that trees are less correlated with each other. By aggregating the predictions of all trees, the model achieves more stable and reliable results.

2. Decision Tree: Decision Tree (DT) is a nonparametric supervised algorithm used for both classification and regression tasks. It is easy to interpret, as it creates simple rules from labeled data, and can handle both numerical and categorical features. DTs perform well even with noisy data, but they have limitations, such as sensitivity to small changes in the data and a tendency to overfit. Additionally, they cannot naturally handle missing values, which can affect their reliability.

3. K-Nearest Neighbours: K-Nearest Neighbours (KNN) is a nonparametric supervised algorithm used for classification and regression by assigning a data point to the class of its nearest neighbours. It relies on the similarity of features to make predictions. The parameter k determines how many neighbours are considered in the majority vote. Different distance measures, like Euclidean, Manhattan, or Minkowski, can be used to calculate closeness between points.

4. Support Vector Machine (SVM): Support Vector Machine (SVM) is a supervised learning method used for both classification and regression tasks. It identifies an optimal decision boundary, known as a hyperplane, that maximizes the margin between different classes. SVMs perform well when the dataset has fewer samples than features and can employ various kernel functions, like linear or RBF, to capture complex patterns. This flexibility allows SVMs to model intricate decision boundaries effectively.

5.Artificial Neural Network: An Artificial Neural Network (ANN) is a supervised learning model inspired by the human brain, composed of interconnected units called neurons. It consists of input, output, and multiple hidden layers, where each connection has an associated weight. Inputs are processed through neurons using weighted sums and activation functions to determine neuron activation. Activated signals are then propagated forward through the network in a process called feed-forward propagation.

IV SYSTEM DEVELOPMENT

A. DATASET USED

This dataset has been developed to support NLP-driven diagnostic analysis within Hospital Network Information Management Systems (HNIMS). It represents simulated electronic health records (EHRs), incorporating digitized medical reports, text extracted through optical character recognition (OCR), and essential clinical indicators associated with sleep disorders.

Dataset link:
https://www.kaggle.com/datasets/ziya07/sleep-disorder-diagnostic-dataset

B. IMPLEMENTED ALGORITHMS

- XGBoost: XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm based on sequential decision trees, where each new tree aims to reduce the errors of the previous ones. It incorporates regularization to prevent overfitting, making it highly efficient, scalable, and popular in practical applications and data science competitions.
- Gradient Boosting: Gradient Boosting is an ensemble method that sequentially builds weak models, typically decision trees, with each new model correcting the errors of the previous ones. By optimizing a loss function through gradient descent, it combines these models to deliver increasingly accurate predictions.
- Logistic Regression: Logistic Regression is a supervised algorithm for binary classification that predicts the probability of an outcome using a logistic (sigmoid) function on input features. It outputs values between 0 and 1, with class labels assigned based on a probability threshold.

C. SYSTEM ARCHITECTURE

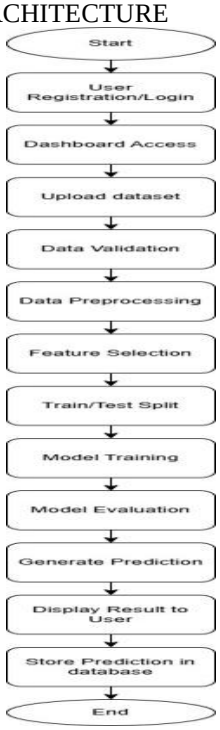

Fig.1. Block Diagram for Sleep Disorder Prediction

Explanation

1.Start
The system begins when the user opens the web application in browser.

2.User Registration/ Login
New users can sign up, and existing users can log in. Authentication is handled by Django's built-in User model, ensuring secure access.

3.Dashboard access
After login, the user is redirected to the dashboard. The dashboard options are: Upload Dataset, View Previous Results, and Logout.

4.Upload Dataset
The user uploads a dataset in CSV format containing sleep-related parameters.

5.Data validation
The system checks for the correct file format, the presence of required columns, and the absence of corrupted data. If the file is invalid, an error message is displayed.

6.Data Preprocessing
Handle missing values, encode categorical variables, normalize numerical values, and remove duplicates.

7.Feature Selection
Relevant features are selected for model training, such as sleep duration, age, and stress level, while irrelevant attributes are removed.

8.Train/Test Split
The dataset is divided into training data (80%) and testing data (20%) to ensure unbiased evaluation.

9.Model Training
The model is trained using the following algorithms: Random Forest, Decision Tree, K-Nearest Neighbours, Support Vector Machine (SVM), Multilayer Perceptron (Neural Network), XGBoost, Gradient Boosting, Logistic Regression.

10.Model Evaluation
The model is tested using the test dataset. Evaluation metrics include precision, recall, F1-score, Support, Accuracy. If the accuracy is acceptable, the model proceeds; if not, it is retrained.

11.Generate Prediction
Using the trained model, predict sleep quality and classify it as Good Sleep, Moderate Sleep, or Poor Sleep.

12.Display result to user
The system shows the prediction result, accuracy score, and graph.

13.Store Prediction in database
Predictions are stored in SQLite, including user ID, dataset reference, prediction result, and timestamp, allowing result history tracking.

14.End
The user can upload another dataset, log out, or exit the application.

D. EVALUATION METRICS

- Precision: Indicates how many of the instances predicted as positive are actually positive, reflecting the model's exactness.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall: Shows how well the model identifies all actual positive instances, measuring completeness.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-Score: Combines precision and recall into a single metric, giving a balanced measure of accuracy when classes are imbalanced.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Support: Support represents the total number of actual samples belonging to a specific class.

$$\text{Support} = TP + FN$$

- Accuracy: Measures the overall correctness of a model by showing the proportion of correctly predicted instances out of all predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

V. RESULT

Model Accuracy:

Model	Accuracy
Logistic Regression	0.907
Random Forest	0.947
KNN	0.933
Decision Tree	0.907
SVM (RBF)	0.907
Gradient Boosting	0.947
MLP (Neural Network)	0.920
XGBoost	0.947

Classification Reports:
Logistic Regression:

Class	Precision	Recall	F1-score	Support
Insomnia	0.93	0.87	0.90	15.0
None	0.91	0.95	0.93	42.0
Sleep Apnea	0.88	0.83	0.86	18.0
Macro Avg	0.91	0.88	0.89	75.0
Weighted Avg	0.91	0.91	0.91	75.0

Random Forest:

Class	Precision	Recall	F1-score	Support
Insomnia	0.93	0.87	0.90	15.0
None	0.93	1.00	0.97	42.0
Sleep Apnea	1.00	0.89	0.94	18.0
macro avg	0.95	0.92	0.93	75.0
weighted avg	0.95	0.95	0.95	75.0

KNN:

Class	Precision	Recall	F1-score	Support
Insomnia	0.87	0.87	0.87	15.0
None	0.93	1.00	0.97	42.0
Sleep Apnea	1.00	0.83	0.91	18.0
macro avg	0.93	0.90	0.91	75.0
weighted avg	0.94	0.93	0.93	75.0

Decision Tree:

Class	Precision	Recall	F1-score	Support
Insomnia	0.81	0.87	0.84	15.0
None	0.93	0.93	0.93	42.0
Sleep Apnea	0.94	0.89	0.91	18.0
macro avg	0.89	0.89	0.89	75.0
weighted avg	0.91	0.91	0.91	75.0

SVM (RBF):

Class	Precision	Recall	F1-score	Support
Insomnia	0.76	0.87	0.81	15.0
None	0.93	0.95	0.94	42.0
Sleep Apnea	1.00	0.83	0.91	18.0
Macro Avg	0.90	0.88	0.89	75.0
Weighted Avg	0.91	0.91	0.91	75.0

Gradient Boosting:

Class	Precision	Recall	F1-score	Support
Insomnia	0.93	0.87	0.90	15.0
None	0.93	1.00	0.97	42.0
Sleep Apnea	1.00	0.89	0.94	18.0
Macro Avg	0.95	0.92	0.93	75.0
Weighted Avg	0.95	0.95	0.95	75.0

MLP (Neural Network)

Class	Precision	Recall	F1-score	Support
Insomnia	0.93	0.87	0.90	15.0
None	0.93	0.95	0.94	42.0
Sleep Apnea	0.89	0.89	0.89	18.0
Macro Avg	0.92	0.90	0.91	75.0
Weighted Avg	0.92	0.92	0.92	75.0

XGBoost:

Class	Precision	Recall	F1-score	Support
Insomnia	0.93	0.87	0.90	15.0
None	0.93	1.00	0.97	42.0
Sleep Apnea	1.00	0.89	0.94	18.0
Macro Avg	0.95	0.92	0.93	75.0
Weighted Avg	0.95	0.95	0.95	75.0

VI. CONCLUSION

Various machine learning algorithms were implemented in this study, including Random Forest, K-Nearest Neighbours, Support Vector Machine (RBF), Multilayer Perceptron (Neural Network), Decision Tree, XGBoost, Gradient Boosting, and Logistic Regression. The results indicate that Tree based ensemble models (Random Forest, Gradient Boosting, XGBoost) achieved the highest accuracy of 94.7% and highest weighted F1-score of 0.95. they provided balanced precision and recall across three sleep disorder classes. Among them XGBoost is preferred due to its strong generalization capability and robustness.

ACKNOWLEDGMENT

Thanks to Ms. Rucha Ravindra Galgali for his help and support throughout this project. His guidance and feedback were very important in making this work better.

REFERENCES

[1] T. S. Alshammari, "Applying Machine Learning Algorithms for the Classification of Sleep Disorders," in *IEEE Access*, vol. 12, pp. 36110-36121, 2024, doi: 10.1109/ACCESS.2024.3374408.

[2] Locharla Ravikumar et al., (2025). SLEEP DISORDER PREDICTION USING MACHINE LEARNING. International Journal of Engineering Technology Research & Management (IJETRM), 09(07).

[3] Dr. V. Shanmugapriya et al., “Sleep disorder prediction using machine learning” *International Journal of Scientific and Advanced Research in Technology*, vol. 11, no. 3, 2025.