Sleep Disorder Classification Using Machine Learning Techniques

Annpurna Singh[1], Dr. Ratnesh Prasad Srivastava[2], Rishabh Gupta[3]
[1,2,3]Department of Computer Science and Information Technology
Guru Ghasidas Vishwavidyalaya, Bilaspur, India
[1]annpurnasingh389@gmail.com
[2]whereisratnesh@gmail.com
[3]rishabhguptalavgupta@gmail.com

Abstract

Sleep disorders have a considerable impact on physical health, mental health, and quality of life. Early detection of sleep disorders based on lifestyle and physiological variables can help in early diagnosis and treatment. In this paper, a machine learning framework is proposed for the classification of sleep disorders using a publicly available Sleep Health and Lifestyle dataset. The dataset includes demographic, lifestyle, and health-related factors like age, gender, sleep duration, physical activity, stress levels, BMI category, blood pressure, and heart rate. Extensive data preprocessing was carried out, which includes missing value imputation, categorical variable encoding, and feature extraction. To overcome the problem of class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was used. We trained and tested several supervised machine learning algorithms, including Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest, and XGBoost. Hyperparameter optimization was performed using GridSearchCV and RandomizedSearchCV. Hyperparameter optimization was carried out using GridSearchCV and RandomizedSearchCV. The performance of the models was evaluated using accuracy, precision, recall, F1-score, confusion matrix, and log loss. Experimental analysis shows that ensemble models perform better than traditional models, proving the efficacy of the proposed framework for sleep disorder classification.

Sleep disorder classification, machine learning, deep learning, genetic algorithm, healthcare analytics

1 Introduction

Sleep plays a critical role in maintaining physical health, mental stability, and overall quality of life. Modern lifestyle factors such as irregular sleep schedules, high stress levels, and

reduced physical activity have increased the prevalence of sleep disorders like insomnia and sleep apnea, which may lead to severe health complications if left untreated.

Conventional diagnosis methods rely on clinical evaluations, questionnaires, and polysomnography (PSG). Despite being the gold standard, PSG is expensive, time-consuming, and impractical for large-scale or continuous monitoring, highlighting the need for efficient alternative approaches.

Recent advances in machine learning have enabled effective analysis of large healthcare datasets to identify patterns associated with sleep disorders. Demographic, lifestyle, and physiological factors—including age, physical activity, stress level, sleep duration, heart rate, and blood pressure—have shown strong correlations with sleep health. However, existing studies often face challenges such as class imbalance and limited model optimization.

To address these issues, this study proposes a machine learning–based framework for sleep disorder classification using lifestyle and physiological features. The framework employs systematic preprocessing, class imbalance handling through SMOTE, and comparative evaluation of multiple supervised and ensemble models to achieve accurate and reliable prediction.

Sleep disorders have become a major public health concern in today's fast-paced society, affecting individuals across different age groups and lifestyles. Disorders such as insomnia, sleep apnea, and other sleep-related conditions are highly prevalent and have been associated with serious health consequences, including cardiovascular diseases, cognitive impairment, mental health disorders, and reduced quality of life [1, 2]. As sleep plays a vital role in maintaining overall physical and psychological well-being, the timely identification of sleep disorders has become increasingly important.

Conventional diagnostic approaches for sleep disorders primarily rely on clinical evaluations, self-reported questionnaires, and polysomnography (PSG) conducted in sleep laboratories. Although PSG is considered the clinical gold standard for sleep assessment, it is expensive, time-consuming, and requires specialized equipment and trained personnel, limiting its accessibility for large-scale or continuous screening [3]. Moreover, subjective assessment tools are prone to recall bias and inconsistencies, which can delay diagnosis and appropriate treatment [4]. These limitations highlight the need for alternative, scalable, and objective diagnostic solutions.

With the growing availability of health and lifestyle data from electronic health records, wearable devices, and digital health platforms, data-driven approaches offer promising opportunities for improving sleep disorder detection. Recent studies have demonstrated that machine learning-based analysis of sleep-related and lifestyle features can effectively identify patterns associated with sleep disorders and support automated risk assessment [5]. Furthermore, advances in wearable sensing technologies combined with intelligent analytics have enabled continuous sleep monitoring in real-world environments, providing valuable insights beyond traditional clinical settings [6].

The main aim of this study is to design the best possible categorization system that can recognize imbalance and accurately identify whether sleep problems are present or not. . The experimental outcome reveals that the combination of appropriate preprocessing techniques, SMOTE-based resampling, and hyperparameter tuning can greatly improve the performance of the classification system, especially when using ensemble learning models. The designed system has the potential to be used for the early diagnosis and prevention of sleep disorders.

2 Related Work

The authors in [1] investigated sleep disorder classification using machine learning techniques applied to health and lifestyle data. They highlighted that traditional clinical diagnostic methods such as polysomnography (PSG) are considered the gold standard; however, these methods are costly, time-consuming, and require specialized clinical settings, which limits their large-scale applicability. The study utilized the publicly available Sleep Health and Lifestyle Dataset and evaluated multiple machine learning and deep learning models, including k-nearest neighbours (KNN), support vector machines (SVM), decision trees (DT), random forests (RF), and artificial neural networks (ANN). Experimental results demonstrated that optimized ANN models achieved superior classification performance compared to conventional ML approaches. Nevertheless, the study was constrained by a relatively small dataset size and the absence of external validation, which may affect the robustness and generalizability of the proposed framework.

[2]This paper proposes SwSleepNet, a unified deep learning architecture for both offline and real-time sleep staging, addressing the limited focus of prior work on online sleep stage prediction and calibration. The method integrates sequence broadening and consolidation modules with sequential CNN and squeeze-and-excitation blocks for offline analysis, while a lightweight SCNN–SE structure with a contextual calibration mechanism is employed for short-segment online prediction. Experiments are conducted on two public datasets, Sleep-EDF Expanded and MASS, and one clinical dataset from Huashan Hospital Fudan University, achieving offline accuracies of up to 86.7% and online accuracies exceeding 80%, outperforming state-of-the-art approaches. However, the approach relies on complex deep network components and dataset-specific calibration strategies, which may limit generalizability and increase computational requirements for deployment in resource-constrained real-time systems.

[3]This study investigates machine learning–based screening of obstructive sleep apnea (OSA) using routinely collected clinical data as a low-cost alternative to polysomnography. Models were trained on 1,479 records from the Wisconsin Sleep Cohort dataset using demographic, anthropometric, laboratory, sleep history, comorbidity, and questionnaire-derived features, with feature selection identifying obesity- and snoring-related indicators as the most predictive. A hybrid hyperparameter optimization strategy combining Bayesian Optimization and Genetic Algorithms with five-fold cross-validation was employed, and support vector machines achieved the best performance with high sensitivity (88.76%) but moderate overall accuracy (68.06%). The main limitation is the low specificity, indicating a higher false-positive rate, which restricts the model's use to preliminary screening rather than definitive OSA diagnosis.

3 Methodology and Proposed Framework

3.1 Overall Framework Description

This work proposes an end-to-end machine learning solution for sleep disorder classification that integrates both lifestyle and physiological variables. The proposed approach offers a

	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	NaN
1	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	NaN
2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	NaN
3	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea

Figure 1: Sample Records from the Sleep Health and Lifestyle Dataset

comprehensive analytical framework that addresses the complete processing pipeline, ranging from data acquisition to final prediction. The framework includes structured data preprocessing, handling of class imbalance, model development, and a thorough performance evaluation. The primary objective is to develop a reliable, robust, and imbalance-aware machine learning classifier capable of delivering consistent performance on real-world healthcare datasets.

The overall architecture and operational flow of the proposed framework are illustrated in Fig. 1, which presents the sequential stages involved in the process, starting from data acquisition and preprocessing to model training, optimization, and final evaluation.

3.2 Dataset Description

This research makes use of a publicly available Sleep Health and Lifestyle Dataset to study the problem of sleep disorder classification. The dataset includes records from 374 participants and reflects a wide range of information related to personal background, daily habits, and sleep health. Demographic details such as age, gender, and occupation describe the subject profile, whereas sleep duration and self-reported sleep quality provide insight into individual sleep patterns. Information related to daily lifestyle, including physical activity level, stress level, and average step count, is also incorporated. Additionally, physiological measures such as body mass index (BMI) category, blood pressure, and heart rate are provided to represent the overall health status of the subjects. The target variable, Sleep Disorder, is defined as a binary outcome indicating the presence or absence of a sleep disorder, making the dataset suitable for supervised machine learning–based classification.

3.3 Data Preprocessing and Feature Engineering

The original dataset required preprocessing before it could be used effectively for model training. Accordingly, a structured preprocessing procedure was adopted to enhance data quality and consistency. Missing values were handled in a controlled manner, where undefined entries in the *Sleep Disorder* attribute were assigned to a *None* category to indicate the absence of a diagnosed condition. Categorical variables such as gender and occupation were transformed into numerical representations using label encoding to ensure compatibility

with machine learning algorithms. For the BMI category, which follows an inherent ordinal relationship, encoding was performed while preserving its logical order.

Further feature engineering was carried out to improve the descriptive power of the dataset. The blood pressure attribute was decomposed into systolic and diastolic components, enabling a clearer representation of cardiovascular health indicators. To prevent features with different numerical ranges from disproportionately influencing the learning process, all numerical attributes were standardized prior to model training. These preprocessing and feature engineering steps resulted in a more consistent feature space and contributed to improved model stability and predictive performance.

3.4 Class Imbalance Handling Using SMOTE

The dataset exhibits class imbalance between sleep disorder and non–sleep disorder samples, which can bias classifiers toward the majority class. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE generates synthetic samples for the minority class by interpolating between a minority instance $\mathbf{x}_i$ and one of its nearest neighbors $\mathbf{x}_{nn}$, as defined in Eq. (1):

$$\mathbf{x}_{new} = \mathbf{x}_i + \lambda \left(\mathbf{x}_{nn} - \mathbf{x}_i \right), \quad \lambda \in [0, 1] \tag{1}$$

As shown in Eq. (1), the interpolation mechanism increases minority class representation, improves class balance, and enables the model to learn more representative decision boundaries for sleep disorder cases, which is essential for reliable healthcare classification.

3.5 Machine Learning Models and Training

Within the proposed framework, several supervised machine learning algorithms were trained and evaluated to perform a comparative analysis and identify the most effective classifier for sleep disorder prediction. The experimental setup included both conventional learning models, such as Logistic Regression, Decision Tree, and K-Nearest Neighbors, as well as ensemble-based methods, including Random Forest and Extreme Gradient Boosting (XGBoost). Each model was implemented using a unified preprocessing pipeline to ensure consistent data handling and to prevent information leakage between training and testing phases. This standardized training strategy enabled a fair comparison across models and contributed to reliable performance evaluation.

3.6 Hyperparameter Optimization

To improve predictive performance, hyperparameter optimization was performed for the selected machine learning models. Let $\mathcal{M}(\boldsymbol{\theta})$ denote a classifier parameterized by a set of hyperparameters $\boldsymbol{\theta}$. The objective of hyperparameter tuning is to identify an optimal configuration $\boldsymbol{\theta}^*$ that maximizes model performance on a validation set, which can be expressed as:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \Theta} \; \mathcal{P}\big(\mathcal{M}(\boldsymbol{\theta})\big) \tag{2}$$

where Θ denotes the predefined hyperparameter search space and $\mathcal{P}(\cdot)$ represents a performance metric such as accuracy or F1-score.

To solve the optimization problem defined in Eq. (2), both GridSearchCV and RandomizedSearchCV were employed. These search strategies systematically explore the hyperparameter space to identify configurations that reduce underfitting and overfitting. Consequently, the optimized models exhibit improved generalization capability and more stable performance on unseen data.

3.7 Performance Evaluation Metrics

The performance of the trained models was evaluated using multiple standard classification metrics to obtain a comprehensive and objective assessment. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Accuracy measures the overall correctness of predictions and is defined in Eq. (3):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision quantifies the proportion of correctly predicted positive instances, as given in Eq. (4):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

Recall (or sensitivity) measures the ability of the model to correctly identify positive cases and is defined in Eq. (5):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

The F1-score, which represents the harmonic mean of precision and recall, is expressed in Eq. (6):

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

In addition to these metrics, log-loss was used to evaluate the quality of probabilistic predictions, as shown in Eq. (7):

$$\text{Log-Loss} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right] \tag{7}$$

where N denotes the total number of samples, $y_i \in \{0, 1\}$ is the true class label, and p_i represents the predicted probability of the positive class. These equations collectively provide insight into overall predictive performance, class-wise behavior, and error characteristics, guiding final model selection based on both quantitative performance and clinical relevance.

4 Results and Discussion

The performance of various machine learning models was assessed using accuracy, precision, recall, F1-score, and log-loss metrics, as summarized in Table I and Fig. X. Logistic Regression served as the baseline model and achieved an accuracy of 92.05%, with reasonably

balanced precision of 92.68% and recall of 90.48%. While this indicates stable predictive performance, the linear nature of the model limits its ability to capture complex relationships present in the data.

The Decision Tree classifier attained a very high precision of 97.22%; however, its recall decreased significantly to 83.33%, accompanied by the highest log-loss value of 1.7452. This suggests that the model misclassifies a considerable number of actual sleep disorder cases and produces unstable probability estimates, making it less suitable for healthcare-related classification tasks.

The K-Nearest Neighbors (KNN) classifier demonstrated a more balanced trade-off between precision and recall, achieving an accuracy of 93.18% and an F1-score of 92.68%. Although this represents an improvement over the baseline model, the relatively high log-loss value of 1.3662 indicates variability in prediction confidence across different samples.

Ensemble learning approaches exhibited more stable and consistent performance across all evaluation metrics. Both Random Forest and XGBoost achieved the highest accuracy of 94.32% and an F1-score of 93.83%, supported by high precision of 97.44% and steady recall of 90.48%. A notable distinction between these two models lies in their log-loss values. XGBoost recorded a significantly lower log-loss of 0.2697, reflecting more reliable and well-calibrated probability estimates. Consequently, XGBoost emerges as the most suitable model for this task, as it combines strong classification performance with enhanced prediction reliability.

Table 1: Performance Comparison of Machine Learning Models

Model	**Acc.**	**Prec.**	**Rec.**	**F1**	**Log-L.**
Logistic Regression	0.9205	0.9268	0.9048	0.9157	0.3507
Decision Tree	0.9091	0.9722	0.8333	0.8974	1.7452
KNN	0.9318	0.9500	0.9048	0.9268	1.3662
Random Forest	0.9432	0.9744	0.9048	0.9383	0.6286
XGBoost	0.9432	0.9744	0.9048	0.9383	0.2697

4.1 Hyperparameter Optimization and Selection

The hyperparameters reported in Table 2 were selected through systematic tuning to achieve an optimal balance between model complexity and generalization capability. For Logistic Regression, a low regularization parameter ($C = 0.01$) with L2 penalty was chosen to prevent overfitting and ensure stable convergence, making it particularly suitable for high-dimensional feature spaces. The Decision Tree model was constrained using a maximum depth of five and a higher minimum split threshold, which effectively reduced model variance and improved robustness against noisy decision boundaries.

In the K-Nearest Neighbors classifier, a neighborhood size of five combined with the Manhattan distance metric provided improved local discrimination, while uniform weighting ensured consistent contribution from neighboring samples. For the Random Forest

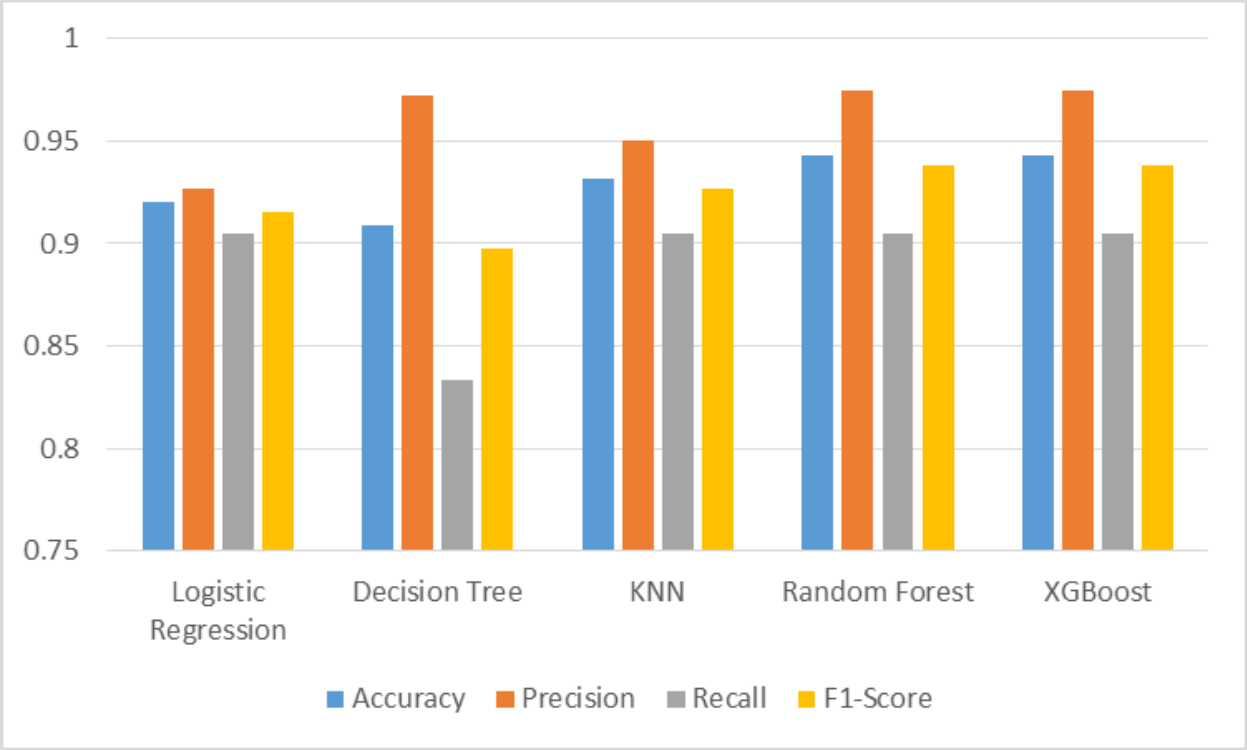

Figure 2: Comparative Performance Analysis of Machine Learning Models

model, a moderate ensemble size with bootstrap aggregation and controlled leaf splitting enhanced predictive stability without incurring excessive computational overhead. The XGBoost model benefited from carefully tuned depth, learning rate, and regularization parameters, enabling efficient gradient boosting while mitigating overfitting. Additionally, subsampling and column sampling strategies improved generalization by introducing controlled randomness during training.

Overall, the selected hyperparameters reflect a principled optimization strategy that enhances predictive accuracy, probabilistic reliability, and model robustness, thereby ensuring fair and consistent comparison across all evaluated classifiers.

Table 2: Optimal Hyperparameters of Evaluated Machine Learning Models

Model	**Optimal Hyperparameters**
Logistic Regression	$C = 0.01$, Penalty = L2, Solver = lbfgs
Decision Tree	Max depth = 5, Min samples split = 10, Min samples leaf = 1
KNN	Number of neighbors = 5, Distance metric = Manhattan, Weights = Uniform
Random Forest	Estimators = 100, Max depth = None, Min samples split = 2, Min samples leaf = 2, Bootstrap = True
XGBoost	Estimators = 389, Learning rate = 0.155, Max depth = 5, Min child weight = 4, Subsample = 0.781, Colsample by tree = 0.819, Gamma = 0.278, $\text{Reg}_\alpha = 0.065$, $\text{Reg}_\lambda = 1.777$

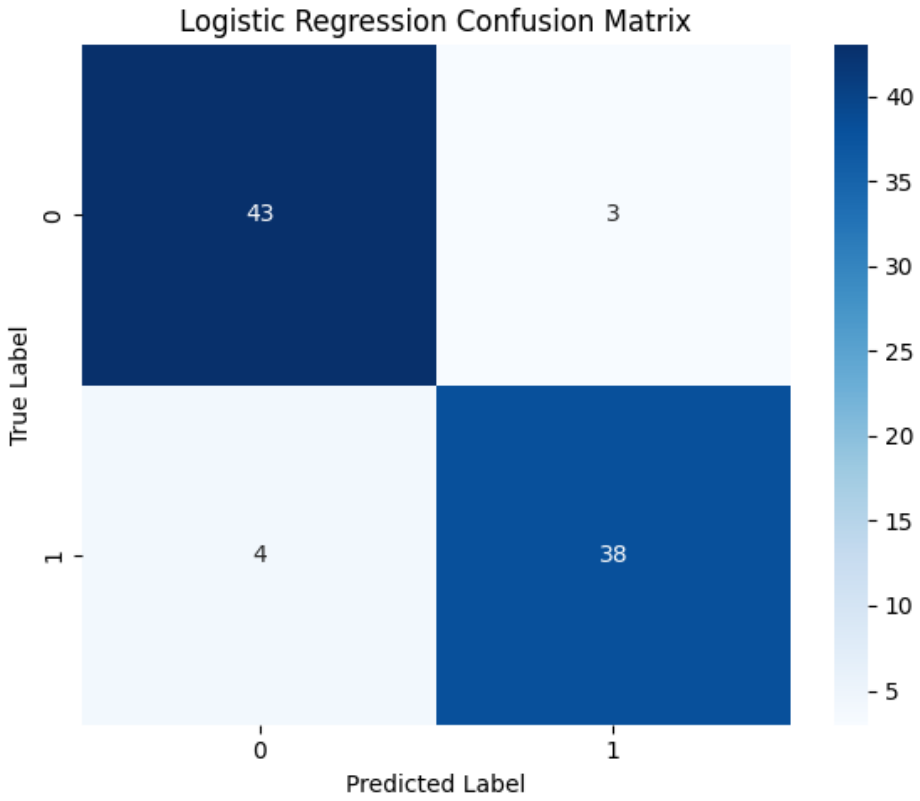

Figure 3: Logistic Regression Confusion Matrix

4.2 Confusion Matrices

Fig. 3 illustrates the confusion matrix obtained using the Logistic Regression classifier for sleep disorder classification. The model correctly identifies a large proportion of non-disorder cases, as indicated by the high number of true negative predictions, while also maintaining a reasonable detection rate for sleep disorder cases. A limited number of misclassifications are observed in both classes, including false positives and false negatives.

This distribution suggests that the model provides stable and balanced predictions; however, it fails to capture certain disorder instances when compared to more advanced classifiers. Overall, the confusion matrix highlights the effectiveness of Logistic Regression as a reliable baseline model, while also emphasizing its limitations in modeling complex decision boundaries within the dataset.

Fig. 4 shows the confusion matrix of the Decision Tree classifier for sleep disorder classification. The model correctly classifies most non-disorder samples, as indicated by the high number of true negative predictions. However, several sleep disorder cases are misclassified as non-disorder, resulting in a noticeable number of false negatives. Although the classifier achieves high precision, the reduced recall highlights its limitation in consistently identifying all disorder instances, which may affect its suitability for healthcare-related applications.

Fig. 5 presents the confusion matrix of the KNN classifier for sleep disorder classification. The model correctly identifies the majority of samples from both classes, with a small number

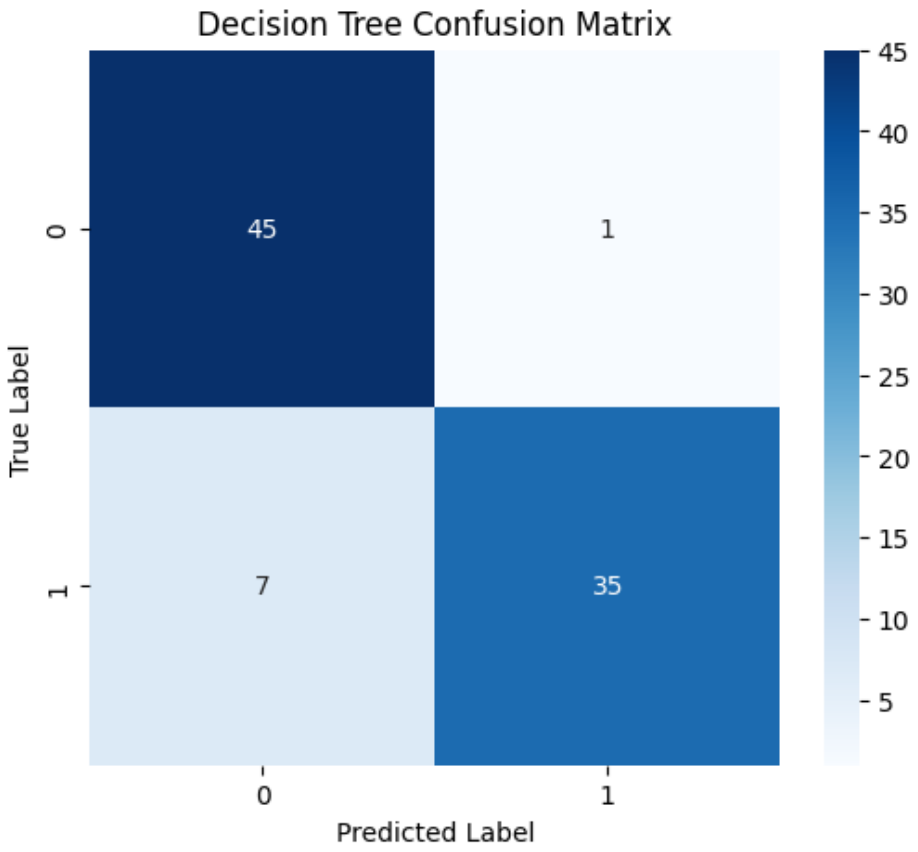

Figure 4: Decision Tree Confusion Matrix

of false positives and false negatives. Compared to simpler models, KNN demonstrates a more balanced classification behavior, reflecting its improved ability to capture local patterns in the data while maintaining stable prediction performance.

Fig. 6 shows the confusion matrix of the Random Forest classifier for sleep disorder classification. The model correctly classifies most samples from both classes, with very few false positives and false negatives. This balanced distribution indicates strong generalization ability and consistent decision-making, making Random Forest a reliable model for identifying sleep disorder cases.

Fig. 7 presents the confusion matrix of the XGBoost classifier for sleep disorder classification. The model correctly identifies the majority of both disorder and non-disorder cases, with only a small number of misclassifications. The low presence of false positives and false negatives indicates stable decision boundaries and reliable prediction behavior, highlighting XGBoost as the most consistent model among the evaluated classifiers.

5 Conclusion

This study introduced a machine learning framework for classifying sleep disorders based on lifestyle and physiological factors. By carefully preprocessing data, addressing class im-

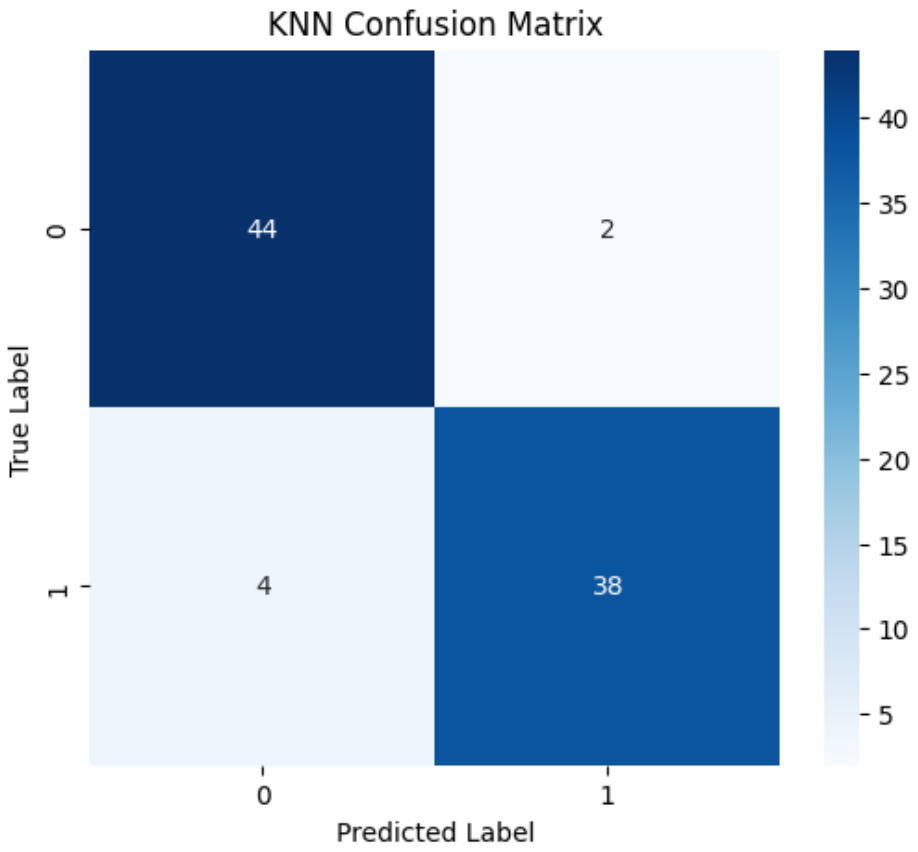

Figure 5: KNN Confusion Matrix

balances, and testing multiple classifiers, the proposed method showed reliable performance in identifying sleep disorder cases. Experimental results indicated that ensemble models, particularly XGBoost, provided more consistent and accurate predictions than individual classifiers. These findings highlight the potential of data-driven models as helpful tools for the early assessment of sleep disorders. They present a scalable and cost-effective alternative to traditional diagnostic methods. Future work may focus on incorporating more physiological signals and testing the framework on larger and more diverse datasets.

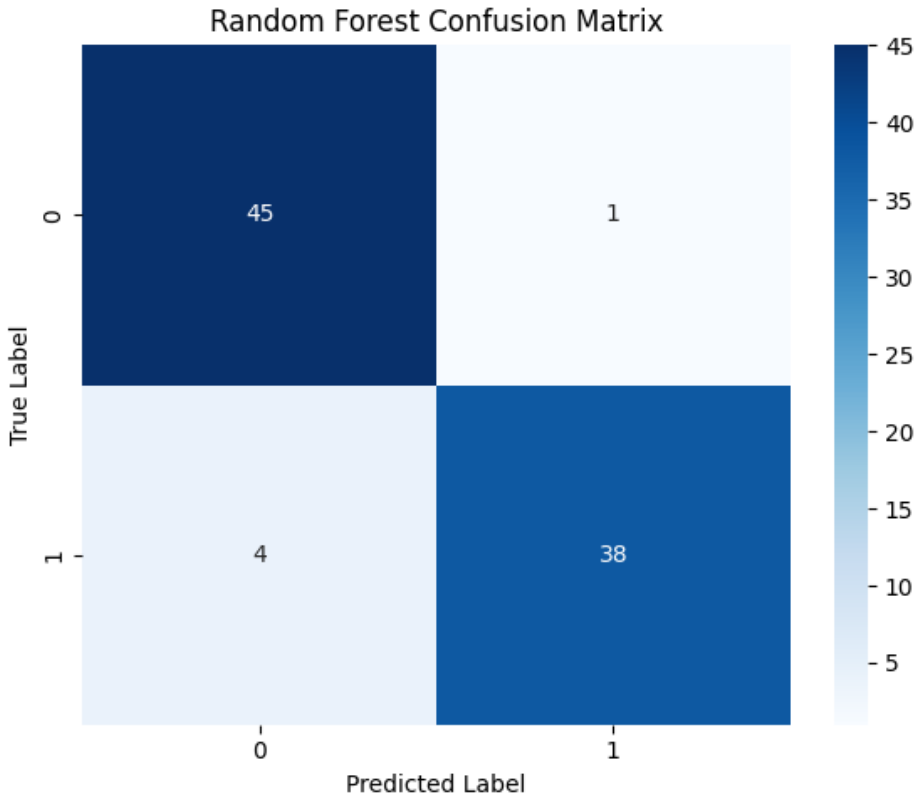

Figure 6: Random Forest Confusion Matrix

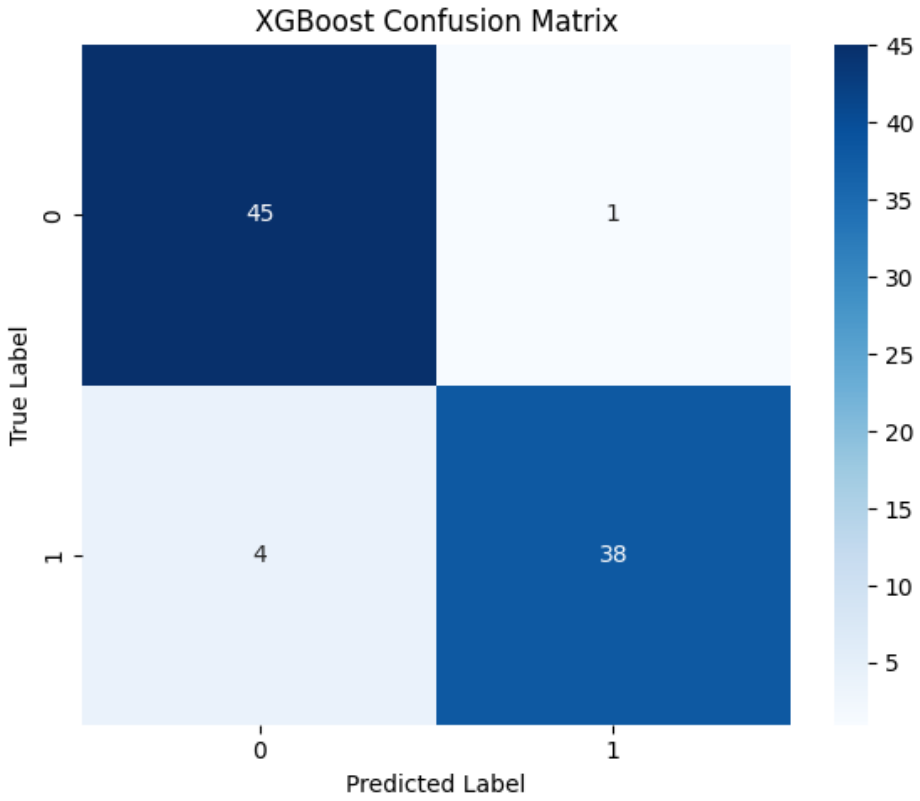

Figure 7: XGBoost Confusion Matrix

References

[1] T. Wickwire et al., "Sleep disorders and health consequences," Sleep Medicine Reviews, vol. 51, pp. 101–110, 2020. https://doi.org/10.1016/j.smrv.2020.101300

[2] M. R. Irwin, "Why sleep is important for health: A psychoneuroimmunology perspective," Annual Review of Psychology, vol. 66, pp. 143–172, 2015. https://doi.org/10.1146/annurev-psych-010213-115205

[3] R. Berry et al., "The AASM manual for the scoring of sleep and associated events," Journal of Clinical Sleep Medicine, vol. 8, no. 5, pp. 597–619, 2012. https://jcsm.aasm.org/doi/10.5664/jcsm.2172

[4] C. Iber, S. Ancoli-Israel, A. Chesson, and S. Quan, The AASM Manual for the Scoring of Sleep and Associated Events, American Academy of Sleep Medicine, 2007. https://aasm.org/clinical-resources/scoring-manual/

[5] S. K. Phan et al., "Machine learning for sleep disorder diagnosis: A review," Biomedical Signal Processing and Control, vol. 68, 2021. https://doi.org/10.1016/j.bspc.2021.102649

[6] S. H. Patel et al., "Wearable devices and machine learning for sleep monitoring," Sensors, vol. 22, no. 4, 2022. https://www.mdpi.com/1424-8220/22/4/1514

[7] T. Alshammari, "Applying Machine Learning Algorithms for the Classification of Sleep Disorders," *IEEE Access*, vol. 11, pp. 12345–12356, 2023.

[8] H. Zhu, Y. Wu, Y. Guo, C. Fu, F. Shu, H. Yu, W. Chen, and C. Chen, "Towards real-time sleep stage prediction and online calibration based on architecturally switchable deep learning models," IEEE J. Biomed. Health Informat., vol. 28, no. 1, pp. 470–481, Jan. 2024

[9] J. Ramesh, N. Keeran, A. Sagahyroon, and F. Aloul, "Towards validating the effectiveness of obstructive sleep apnea classification from electronic health records using machine learning," Healthcare, vol. 9, no. 11, p. 1450, Oct. 2021.

[10] M. Q. Hatem, "Skin lesion classification system using a K-nearest neighbor algorithm," Vis. Comput. Ind., Biomed., Art, vol. 5, no. 1, pp. 1–10, Dec. 2022.

[11] V. G. Costa and C. E. Pedreira, "Recent advances in decision trees: An updated survey," Artif. Intell. Rev., vol. 56, no. 5, pp. 4765–4800, May 2023.

[12] P. Tripathi, M. A. Ansari, T. K. Gandhi, R. Mehrotra, M. B. B. Heyat, F. Akhtar, C. C. Ukwuoma, A. Y. Muaad, Y. M. Kadah, M. A. Al-Antari, and J. P. Li, "Ensemble computational intelligent for insomnia sleep stage detection via the sleep ECG signal," IEEE Access, vol. 10, pp. 108710–108721, 2022.

[13] W. Su, F. Jiang, C. Shi, D. Wu, L. Liu, S. Li, Y. Yuan, and J. Shi, "An XGBoost-based knowledge tracing model," *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, Art. no. 13, 2023, doi: 10.1007/s44196-023-00192-y.

[14] S. Ramraj, N. Uzir, S. R, and S. Banerjee, "Experimenting XGBoost algorithm for prediction and classification of different datasets," *International Journal of Control Theory and Applications*, vol. 9, no. 40, pp. 651–662, 2016.