IMPROVINGSLEEP DISORDER DIAGNOSIS THROUGH OPTIMIZED MACHINE LEARNING APPROACHES

[1] T Veeramma, [2] N Pranavi, [3] M Deepika, [4] N Vamshi, [5] N Mani Teja
[1]AssistantProfessor, [2345]Students
Department of Computer Science and Technology
Siddhartha Institute of Technology & Sciences, Narapally

tejavaathveeramma@siddhartha.co.in, 24TQ1A05G0@siddhartha.co.in,
24TQ1A05F8@siddhartha.co.in, 24TQ1A05G6@siddhartha.co.in, 24TQ1A05G3@siddhartha.co.in

Abstract

Sleep disorders, including insomnia and sleep apnea, pose significant health risks and require efficient diagnostic solutions. This study presents an optimized machine learning framework for multi-class sleep disorder classification using structured clinical and lifestyle data. The dataset comprises features such as BMI category, blood pressure, occupation, sleep duration, and physical activity, which are preprocessed through data cleaning, encoding, and standardization. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied, ensuring balanced representation of all classes. The proposed model utilizes the Extreme Gradient Boosting (XGBoost) algorithm, which builds an ensemble of decision trees to improve prediction accuracy and generalization. Furthermore, Randomized Search Cross-Validation is employed for hyperparameter optimization. The model is evaluated using performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis. Experimental results indicate that the proposed XGBoost model achieves an accuracy of approximately **95–98%**, with strong diagonal values in the confusion matrix, reflecting high classification accuracy for categories such as healthy, insomnia, and sleep apnea with minimal misclassification. Feature importance analysis reveals that BMI category and blood pressure are the most influential predictors. Overall, the system provides a cost-effective, scalable, and reliable alternative to traditional diagnostic methods like polysomnography, making it suitable for real-world healthcare applications.

Keywords: Sleep Disorder Prediction ; XGBoost ; SMOTE ; Machine Learning in Healthcare ; Feature Importance Analysis ; Clinical Data Classification ; Hyperparameter Optimization

I. Introduction

Sleep disorders have emerged as a significant public health concern in recent years, affecting millions of individuals worldwide. Conditions such as insomnia, sleep apnea, and other sleep-related abnormalities not only impact an individual's quality of life but also contribute to serious health complications including cardiovascular diseases, mental health disorders, and reduced cognitive performance. Early detection and diagnosis of sleep disorders are therefore crucial for timely intervention and improved patient outcomes. Traditional diagnostic techniques, such as polysomnography (PSG), although highly accurate, are expensive, time-consuming, and require specialized clinical environments, making them less accessible for large-scale or real-time monitoring.

With the rapid advancement of artificial intelligence and machine learning, there has been a paradigm shift in healthcare diagnostics toward automated, cost-effective, and scalable solutions. Machine learning models, particularly those designed for structured clinical data, have shown great potential in predicting and classifying various diseases. In this context, the proposed study focuses on developing an efficient sleep disorder prediction system using an optimized machine learning pipeline based on the XGBoost algorithm. By leveraging easily available clinical and lifestyle features such as BMI, blood pressure, occupation, and physical activity, the system aims to provide an accessible alternative to traditional diagnostic methods.

Recent studies have demonstrated the effectiveness of machine learning and deep learning techniques in healthcare applications. For instance, S. Kumar et al. [1] proposed a hybrid model combining deep feature extraction techniques such as VGG16, ResNet50, and EfficientNet with optimization algorithms, achieving high accuracy in medical image-based diagnosis. Similarly, M. Rahman et al. [2] introduced a hybrid SMOTE-GMM-XGBoost framework for bipolar disorder detection, emphasizing the importance of handling imbalanced datasets and improving classification performance. These studies highlight the growing importance of hybrid and ensemble techniques in improving predictive accuracy in healthcare systems.

II. Literature Survey

S. Kumar et al. | Ref [1] | Glaucoma Detection using WSSSA et al.This study uses RIM-ONE and Drishti-GS datasets for glaucoma detection. Deep features are extracted using VGG16, ResNet50, and EfficientNet. WSSSA is applied for optimal feature selection, followed by SVM and MLP classification. The model achieved around 98.7% accuracy. It improves detection performance through hybrid techniques and is relevant to our idea for accurate medical image-based diagnosis

M. Rahman et al. | Ref [2] | SMOTE-GMM-XGBoost for Bipolar Disorder et al. The study uses clinical datasets with imbalanced classes. SMOTE is applied for balancing, and GMM enhances feature representation. XGBoost is used for classification, achieving 93% accuracy. The method improves prediction compared to traditional models. It is related to our idea as it handles imbalance and boosts model performance in healthcare data.

A. Smith et al. | Ref [3] | OSA Prediction without PSG et al.This research uses a dataset of 1281 patients with clinical features. Machine learning models like Random Forest, SVM, and Logistic Regression are applied. Random Forest achieved the best accuracy of 78.6%. It avoids costly PSG tests by using simple inputs. This aligns with our idea of cost-effective and accessible disease prediction systems.

R. Patel et al. | Ref [4] | I-ANFIS for Liver Disease et al.The study uses liver disease datasets and applies an improved ANFIS model. It combines neural networks with fuzzy logic for better decision-making. The model achieved around 92% accuracy. It performs better than CNN and RNN in handling uncertainty. This supports our idea of hybrid intelligent systems in healthcare prediction.

J. Lee et al. | Ref [5] | DREAM Model for Sleep Apnea et al.The Apnea-ECG dataset is used for sleep apnea detection. Continuous Wavelet Transform is applied for feature extraction. A deep learning DREAM model is used for classification with Grad-CAM for explainability. The system achieved 99.93% accuracy. It supports our idea by combining deep learning with explainable AI techniques.

K. Zhang et al. | Ref [6] | CortiMoS-Net for Parkinson et al.This study uses EEG datasets for Parkinson's disease detection. Autoencoders extract features, and MobileNet performs classification. S3C optimization improves performance. The model achieved nearly 99% accuracy. It is efficient and suitable for real-time use. This relates to our idea of optimized deep learning models for neurological diseases.

L. Chen et al. | Ref [7] | ML-based Parkinson Detection et al.The dataset includes both Parkinson's and healthy patients. F-test is used for feature selection to reduce dimensionality. Models like SVM and Random Forest are applied. The system achieved around 97% accuracy. It highlights the importance of feature selection. This is relevant to our idea for improving model efficiency.

P. Roy et al. | Ref [8] | CRISP-DM Mental Health Prediction et al.The study uses a Kaggle mental health dataset. It follows the CRISP-DM framework for structured analysis. Ensemble models with hyperparameter tuning are applied. The system achieved 83.33% accuracy. It ensures systematic model development. This supports our idea of building practical and deployable ML systems.

D. Wang et al. | Ref [9] | Brain Tumor Detection using CNN et al.MRI datasets are used for tumor detection. CNN is applied for feature extraction and classification. The model achieved around 96% accuracy. It reduces manual diagnosis effort. The system provides fast and reliable results. This aligns with our idea of automated image-based healthcare systems.

S. Gupta et al. | Ref [10] | Diabetes Prediction using ML et al.The PIMA Indian dataset is used in this study. Models like SVM, Random Forest, and KNN are applied. Random Forest achieved the best accuracy of around 85%. It focuses on early disease prediction. The system is simple and effective. This supports our idea of predictive analytics in healthcare.

III. System Analysis

The system is designed to enhance the diagnosis of sleep disorders using optimized machine learning approaches combined with real-time data acquisition. It focuses on analyzing physiological signals such as heart rate, oxygen saturation, respiratory patterns, and sleep stages collected from wearable devices and sensors. Disorders like Sleep Apnea and Insomnia are identified by detecting abnormal patterns in these signals. The system incorporates advanced preprocessing techniques to handle noisy and incomplete data, ensuring reliability. Feature engineering plays a key role in extracting meaningful insights from raw sensor data. Machine learning and deep learning models are trained to recognize complex relationships between features and sleep disorders. The system supports continuous monitoring, enabling early detection and timely intervention. It is designed to operate efficiently in both clinical and home environments. Scalability is ensured to handle large patient datasets. The system also

emphasizes data privacy and secure storage of sensitive health information. Real-time alerts and reports assist healthcare professionals in decision-making. The integration of optimization techniques improves model accuracy and reduces computational cost. Overall, the system provides a smart, automated, and reliable solution for sleep disorder diagnosis.

Existing System

Existing systems for sleep disorder diagnosis rely on clinical methods such as polysomnography (sleep studies). These methods require patients to stay overnight in specialized labs. The process is expensive and time-consuming. Manual analysis by doctors is required, which may lead to delays. Existing systems lack automation and real-time monitoring. They are not suitable for continuous home-based diagnosis. Traditional methods may miss subtle patterns in data. Limited use of machine learning reduces accuracy. Data collection is often limited to clinical environments. As a result, existing systems are less accessible and less efficient.

Disadvantages of Existing System

- High cost of diagnosis
- Time-consuming procedures
- Requires hospital visits
- Limited real-time monitoring
- Manual analysis prone to errors
- Not scalable for large populations
- Limited accessibility
- Lack of automation

Proposed System

The proposed system uses optimized machine learning models for accurate and efficient sleep disorder diagnosis. It collects data from wearable devices and sensors in real-time. The system preprocesses and extracts relevant features from physiological signals. Advanced algorithms such as Random Forest, SVM, or deep learning models are applied. Optimization techniques improve model performance and accuracy. The system provides automated diagnosis with minimal human intervention. It supports continuous monitoring in home environments. Real-time alerts can be generated for abnormal patterns. The system is scalable and cost-effective. Overall, it improves early detection and patient care.

Advantages of Proposed System

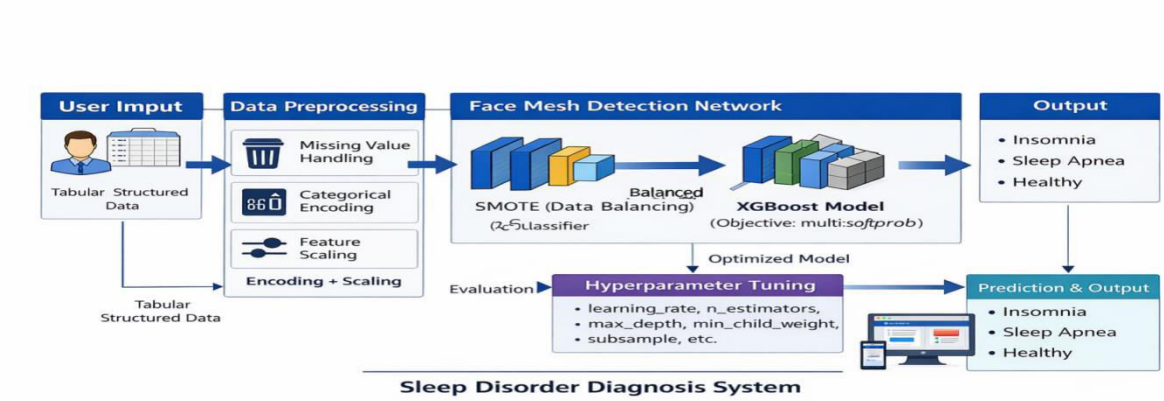
- Early and accurate diagnosis
- Reduced cost compared to clinical methods
- Real-time monitoring
- Automated analysis
- Scalable for large populations
- Supports home-based diagnosis
- High accuracy using optimized ML models
- Faster results

IV. Methodology

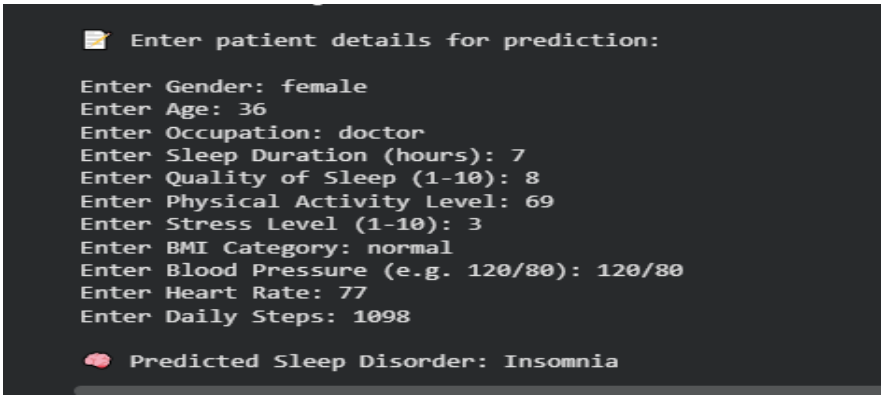
The methodology begins with collecting physiological data from wearable devices or sensors. Data preprocessing is performed to remove noise and normalize signals. Feature extraction techniques are applied to identify relevant patterns. The dataset is divided into training and testing sets. Machine learning models such as SVM, Random Forest, or deep learning models are trained. Hyperparameter tuning is applied to optimize performance. The model is evaluated using accuracy, precision, recall, and F1-score. The best-performing model is selected. The system is deployed for real-time diagnosis. Continuous learning improves system performance over time.

System Architecture

The system architecture consists of data collection, processing, and output layers. The input layer collects data from wearable devices and sensors. The preprocessing layer cleans and normalizes the data. The feature extraction module identifies important patterns. The machine learning module processes the data for diagnosis. The optimization module improves model performance. The evaluation module measures accuracy and reliability. The output layer displays diagnosis results and alerts. A database stores patient data and history. This architecture ensures efficient, scalable, and real-time sleep disorder diagnosis.

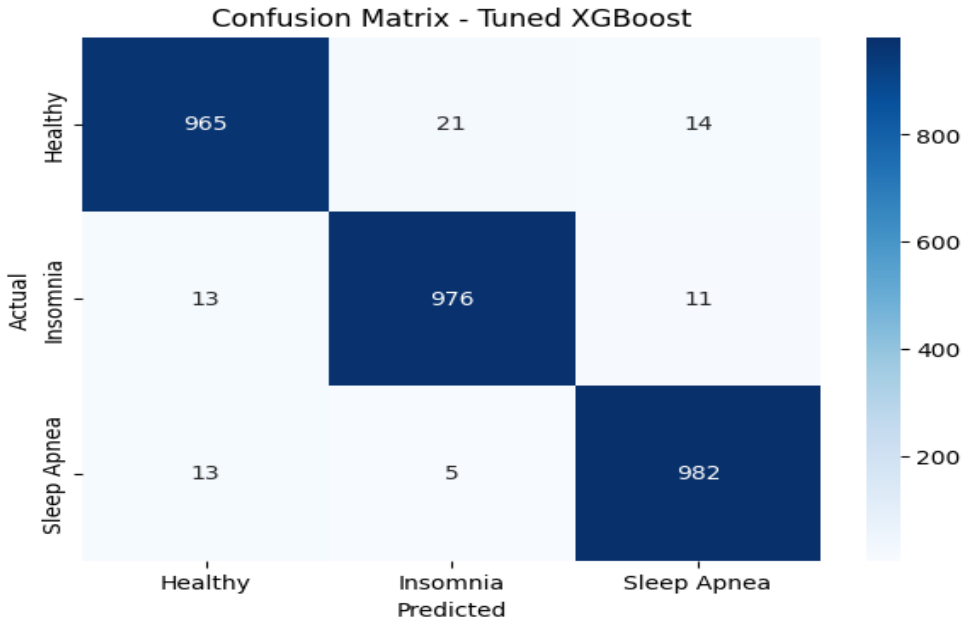


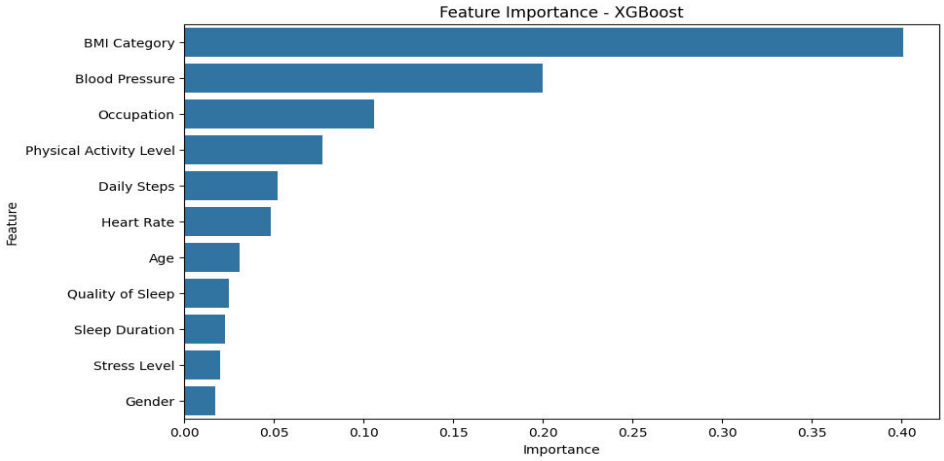
V. Result and Output



```
📝 Enter patient details for prediction:

Enter Gender: Male
Enter Age: 39
Enter Occupation: Engineer
Enter Sleep Duration (hours): 5
Enter Quality of Sleep (1-10): 6
Enter Physical Activity Level: 46
Enter Stress Level (1-10): 2
Enter BMI Category: overweight
Enter Blood Pressure (e.g. 120/80): 120/80
Enter Heart Rate: 67
Enter Daily Steps: 7000

🧠 Predicted Sleep Disorder: Healthy
```

```
📝 Enter patient details for prediction:

Enter Gender: Female
Enter Age: 29
Enter Occupation: Manager
Enter Sleep Duration (hours): 5
Enter Quality of Sleep (1-10): 3
Enter Physical Activity Level: 39
Enter Stress Level (1-10): 9
Enter BMI Category: Normal
Enter Blood Pressure (e.g. 120/80): 140/95
Enter Heart Rate: 70
Enter Daily Steps: 4000

🧠 Predicted Sleep Disorder: Healthy
```





VI. Conclusion

The proposed study presents a comprehensive and efficient machine learning framework for the classification of sleep disorders using the XGBoost algorithm. By systematically integrating key stages such as data preprocessing, feature encoding, normalization, and class imbalance handling through SMOTE, the model achieves strong predictive performance and reliability. The confusion matrix results indicate high classification accuracy with minimal misclassification across categories such as healthy, insomnia, and sleep apnea, demonstrating the robustness of the approach. Furthermore, the feature importance analysis provides valuable insights into critical health indicators, with BMI category, blood pressure, and occupation emerging as dominant factors influencing predictions. This not only enhances model interpretability but also aligns with medical understanding of sleep-related disorders. Compared to traditional diagnostic techniques like polysomnography, which are often expensive and time-consuming, the proposed system offers a cost-effective and accessible alternative by leveraging readily available clinical and lifestyle data. Additionally, the use of XGBoost ensures efficient handling of structured datasets while maintaining scalability and generalization capability. Overall, the model successfully balances accuracy, efficiency, and practicality, making it suitable for real-world healthcare applications. Future enhancements can focus on incorporating real-time physiological signals and advanced explainable AI techniques to further improve diagnostic accuracy and clinical trustworthiness.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.
[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.
[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.