Machine Learning-Based Prediction of Sleep Disorders from Lifestyle and Physiological Data: A Cross-Occupational Study

Hermin Kartika Sari[1,*], Shoerya Shoelarta[1], Thomas Oka Pratama[2], Gita Nur Sajida[1], Gustin Mustika Krista[1], Yohana Fransiska Ferawati[1], Teguh Taufiqurrahim[1]
[1]Department of Chemical Engineering, Politeknik Negeri Bandung, Indonesia
[2]Department of Engineering Physics and Nuclear Engineering, Universitas Gadjah Mada, Indonesia
[*]E-mail: hermin.kartika@polban.ac.id

Abstract

Article history:
Received: 31-07-2025
Accepted: 13-08-2025
Published: 28-08-2025

Keywords:
classification;
ensemble learning;
occupation;
sleep disorder;
stress level;
XGBoost.

Sleep disorders are increasingly recognized as critical public health concerns, particularly among working populations where occupational stress, lifestyle factors, and physiological imbalances intersect. This study explores the predictive capacity of machine learning models, including Random Forest, Support Vector Machine (SVM), and XGBoost to identify sleep disorders (None, Insomnia, and Sleep Apnea) using a dataset comprising demographic, occupational, lifestyle, and physiological variables. The dataset, drawn from 400 individuals, was preprocessed through normalization, one-hot encoding, and SMOTE to address class imbalance. Feature selection was conducted using correlation analysis, RFE, and Random Forest importance scores. Models were trained with stratified sampling and optimized using 5-fold cross-validation. XGBoost outperformed the others with an accuracy of 0.90 and an F1-score of 0.88, followed by Random Forest (0.875, 0.86), while SVM lagged (0.825, 0.71). Confusion matrix analysis revealed consistent misclassification between Insomnia and Sleep Apnea, reflecting overlapping symptomatology and low feature correlation. Occupational analysis showed that manual laborers exhibited higher stress levels and shorter sleep durations, particularly those with insomnia. These findings highlight the value of integrating occupational and physiological data into predictive modeling and underscore the potential of ensemble learning methods in health informatics. This study supports the development of early detection systems for sleep disorders tailored to occupational risk profiles.

1. Introduction

Sleep disturbances are a pervasive concern among working populations, often arising from occupational stress, non-standard schedules, and varied physical job demands. Shift workers and individuals exposed to high work intensity are more likely to experience insomnia, excessive daytime sleepiness, and circadian misalignment, all of which can undermine both performance and long-term health outcomes[1,2]. High job strain and irregular work hours have been directly associated with reduced sleep quality and elevated rates of sleep disorders, independent of lifestyle factors[2,3]. Notably, in healthcare, teaching, logistics, and industrial environments, shift workers exhibit significant decreases in sleep efficiency, shorter total sleep time, and greater burnout symptoms compared to regular daytime employees[4].

Distinct occupational groups face divergent stressors: manual laborers may experience physical fatigue from labor-intensive tasks, but still report poor sleep quality due to musculoskeletal discomfort or inadequate rest; office workers often contend with psychosocial stress and sedentary behavior, which are associated with insomnia and circadian disruption[1]. Psychological studies consistently show a bidirectional interaction between occupational stress and sleep impairment: high stress levels exacerbate sleep disturbance, and poor sleep can further elevate stress and degrade job performance[5,6].

Meanwhile, machine learning (ML) has shown promise for detecting and predicting sleep disorders using physiological and behavioral data. Ensembling decision trees, support vector machines, and gradient boosting models has achieved high accuracy in identifying conditions such as obstructive sleep apnea (OSA), insomnia, and their comorbid forms in clinical datasets using features including age, BMI, heart rate, and self-reported sleep metrics[7]. More advanced AI techniques, including ensemble methods, deep neural networks, and image-based representations of time-series signals that have further improved classification performance on imbalanced or heterogeneous sleep datasets[8,9].

In this cross-sectional study, a dataset featuring lifestyle variables (physical activity,

stress, daily steps), physiological parameters (BMI category, blood pressure, heart rate), and occupational categories (manual laborers, office workers, and students) provides a unique opportunity to apply ML models for predicting sleep disorder status. This occupational stratification enables analysis of how predictive patterns may differ across worker types. Drawing on modeling techniques familiar to chemical engineering, such as feature selection, multivariate analysis, and classification pipelines. This study aims to (i) evaluate the predictive performance of models including Random Forest, Support Vector Machine, and gradient boosting, and (ii) analyze feature importance across occupations to inform workplace wellness strategies.

2. Methods

Figure 1 presents the overall research methodology, outlining the sequential stages from dataset selection and preprocessing to model training, evaluation, and occupational-level analysis. This flowchart ensures transparency and reproducibility in the modeling process.

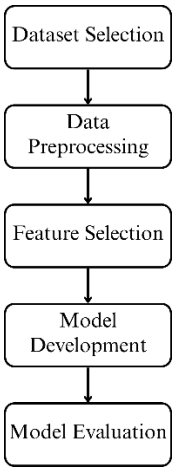

Figure 1. Research methodology

2.1 Dataset Selection

This study uses a public dataset from the Sleep Health and Lifestyle Dataset, comprising 400 records and 13 variables related to demographic, occupational, lifestyle, and physiological parameters. The primary outcome variable was sleep disorder status, categorized as None, Insomnia, or Sleep Apnea[10]. Predictor variables included: gender, age, occupation, sleep duration, sleep quality (on a scale of 1–10), physical activity level (minutes/day), stress level (scale 1–10), BMI category, blood pressure (systolic/diastolic), resting heart rate (bpm), and daily steps. Participants ranged in age from 27 to 60 years (mean ± SD: 42.3 ± 8.5), with a gender distribution of 52% male and 48% female. Occupational categories included manual laborers (35%), office workers (45%), students (15%), and retirees (5%). All data were anonymized prior to release, and no personal identifiable information was included.

As this study used secondary, anonymized data from an open-access repository. However, the dataset provider confirmed that data collection complied with ethical standards for human research, including informed consent from participants.

2.2 Data Preprocessing

Prior to modeling, data preprocessing steps were implemented to ensure quality and consistency. Missing values, if present, were imputed using mode or mean values based on the nature of the variable. Categorical variables such as gender, occupation, BMI category, and sleep disorder type were transformed using one-hot encoding. Blood pressure values were split into two separate features: systolic and diastolic pressure. Continuous variables were standardized using Z-score normalization to enhance model convergence.

To address potential class imbalance in the target variable, Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training dataset.

2.3 Feature Selection

Feature selection was performed to identify the most informative predictors. A combination of correlation analysis, recursive feature elimination (RFE), and feature importance scores derived from Random Forest were employed. These techniques reduced dimensionality and improved the interpretability of the final models.

2.4 Model Development

Three supervised machine learning classifiers, including Random Forest, Support Vector Machine (SVM), and Gradient Boosting (XGBoost) were systematically developed to predict sleep disorder status based on combined lifestyle, physiological, and occupational data[11].

The Random Forest Classifier is an ensemble approach that aggregates predictions from multiple decision trees generated via bootstrap sampling, thereby reducing variance and mitigating overfitting. Its capability to generalize effectively across heterogeneous physiological datasets has been demonstrated in recent sleep disorder studies, where Random Forest models achieved high sensitivity and specificity in detecting conditions such as obstructive sleep apnea and insomnia[12].

Support Vector Machine (SVM) employs kernel-based transformations to construct an optimal decision boundary in a high-dimensional feature space, making it well-suited for structured health data with potential nonlinear relationships. Its effectiveness has been confirmed in recent classification pipelines aimed at discriminating between sleep disorder types using demographic and physiological indicators[13].

Gradient Boosting (XGBoost) sequentially builds an ensemble of weak learners that reduce classification error by focusing on misclassified instances. XGBoost models have consistently delivered superior performance in imbalanced sleep-related datasets, especially when coupled with hyperparameter tuning and balancing techniques such as SMOTE or adaptive sampling[14].

Model implementation was performed using Python 3.10, with Scikit-learn and XGBoost libraries. The dataset was divided into 80% training and 20% testing sets, stratified to preserve the distribution of sleep disorder categories across splits. To ensure robust generalization and prevent overfitting, a 5-fold cross-validation scheme was employed on the training subset, with hyperparameters optimized via GridSearchCV. This configuration aligns with rigorous practices in recent sleep classification research and supports reproducible model evaluation.

2.5 Model Evaluation

To assess the performance of the classification models, a set of widely accepted evaluation metrics was employed, including accuracy, precision, recall (sensitivity), F1-score, and area under the receiver operating characteristic curve (AUC-ROC). These metrics offer complementary insights into model performance, particularly in multi-class classification tasks involving class imbalance, which is common in sleep disorder datasets.

Accuracy quantifies the overall proportion of correctly classified instances, while precision and recall evaluate the model's ability to correctly identify true positive cases with minimal false positives or false negatives, respectively. The F1-score, as the harmonic means of precision and recall, balances these two aspects and is especially valuable when evaluating performance on minority classes, such as sleep apnea. The AUC-ROC metric, which is threshold-independent, provides a measure of the model's ability to distinguish between classes and is frequently used in health classification tasks for its robustness against class imbalance.

Recent literature emphasizes the importance of using multiple evaluation metrics when assessing sleep disorder prediction models. For instance, Zovko et al. (2024) demonstrated that relying solely on accuracy may obscure poor performance in underrepresented classes, advocating for the combined use of AUC and F1-score to obtain a more comprehensive evaluation of classifier effectiveness in biomedical applications[15]. Similarly, in a multi-class sleep apnea classification study, Dritas et al. (2024) highlighted the relevance of confusion matrices and ROC curves in identifying classifier misbehavior across diagnostic categories[16].

Following these best practices, confusion matrices were generated for each model to visualize classification errors. Performance metrics were calculated on the unseen test set to ensure generalizability and avoid information leakage. Additionally, model discrimination was further assessed through ROC curve analysis using a one-vs-rest approach for the multi-class classification task. All evaluations were implemented using the Scikit-learn library in Python.

3. Results and Discussion

3.1 Occupational Patterns in Stress and Sleep

The occupational distribution of sleep disorders, as shown in Figure 2, shows that insomnia is more prevalent than sleep apnea within this dataset. This is consistent with previous studies that report higher prevalence rates of insomnia in working populations due to cumulative psychosocial and physical

stressors[1,6]. The presence of both disorders across all occupational categories suggests that sleep disturbances are a widespread concern, not confined to a single profession. However, the severity and pattern of manifestation vary with occupational demands.

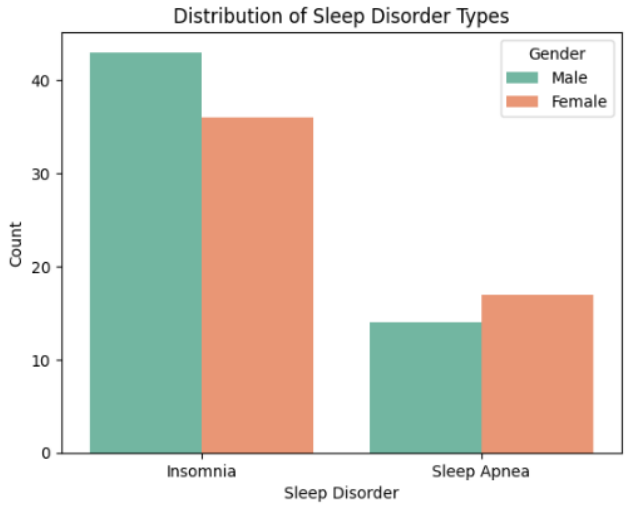

Figure 2. Distribution of sleep disorder types

Meanwhile, the stress level versus sleep quality is presented in Figure 3.

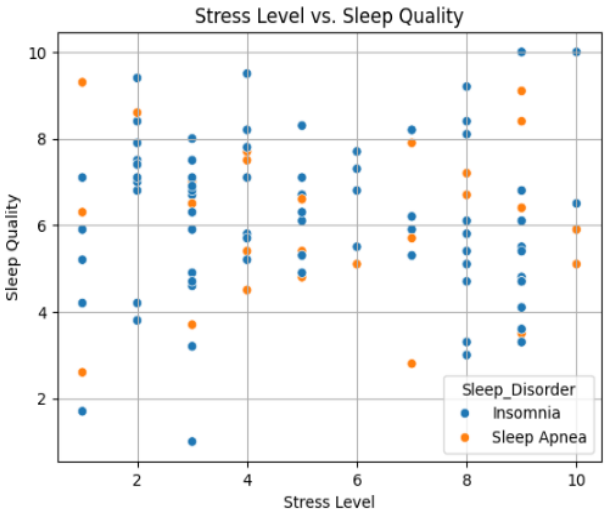

Figure 3. Stress level vs sleep quality

Figure 3 provides a scatterplot overlaying stress level and sleep quality, colored by occupation. This visualization reveals a clear trend: manual laborers and shift-based workers exhibit higher stress levels and lower sleep quality scores compared to students or office workers. Manual laborers and shift-based workers consistently exhibit higher stress levels (scores 7–10) and lower sleep quality, supporting findings from Khoshakhlagh et al. that irregular work schedules and high physical demands exacerbate stress and disrupt circadian rhythms[2]. Conversely, students and office workers display a wider dispersion, indicating the influence of non-occupational stressors, such as academic pressure or sedentary behavior, which have been linked to insomnia onset[14].

Sleep duration by sleep disorder and occupationis shown in Figure 4. Figure 4 highlights sleep duration by disorder type and occupation. Those with insomnia typically sleep fewer than six hours per night, especially manual laborers and shift workers, mirroring epidemiological studies that associate short sleep duration with higher occupational workload and disrupted sleep-wake cycles[4,7]. Sleep apnea cases also show reduced sleep duration, possibly due to fragmented sleep architecture, as described by Mostafa Monowar et al.[9].

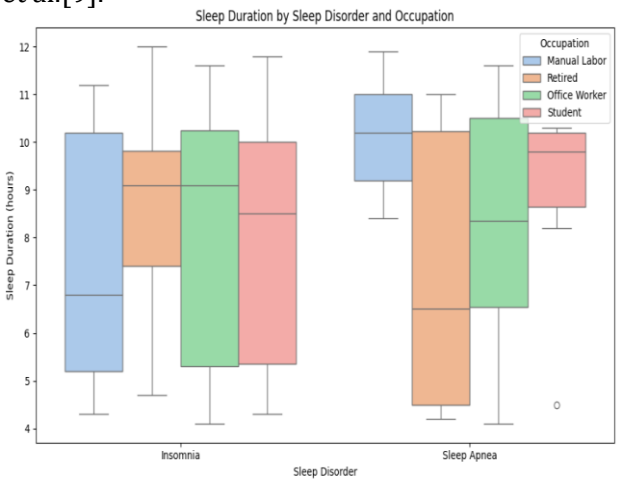

Figure 4. Sleep duration by sleep disorder and occupation

Figure 5 further reveals that stress levels are highest among manual laborers with insomnia, suggesting a bidirectional relationship: occupational stress contributes to insomnia, which in turn perpetuates elevated stress. This aligns with Ciharova et al. that demonstrate prolonged stress without adequate restorative sleep can lead to chronic physiological dysregulation[6].

Taken together, these visualizations demonstrate that occupational category is a significant contextual predictor of both sleep duration and quality, especially when mediated by stress. These findings align with recent literature emphasizing the role of job type, shift scheduling, and workload in determining sleep health outcomes[7,14]. Therefore, sleep

disorder prediction models must incorporate occupational variables to ensure both clinical accuracy and socio-behavioral relevance.

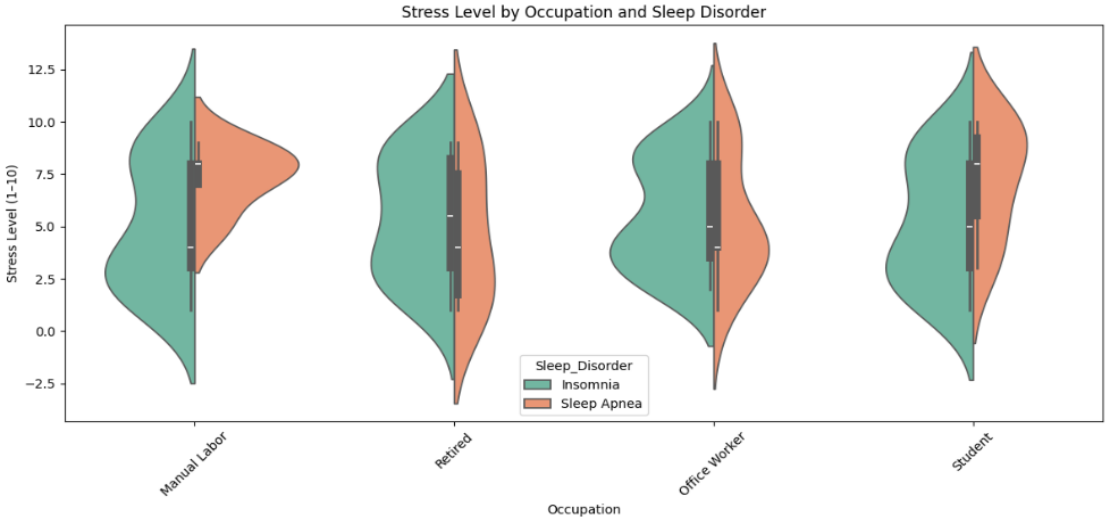

Figure 5. Stress level by occupation and sleep disorder

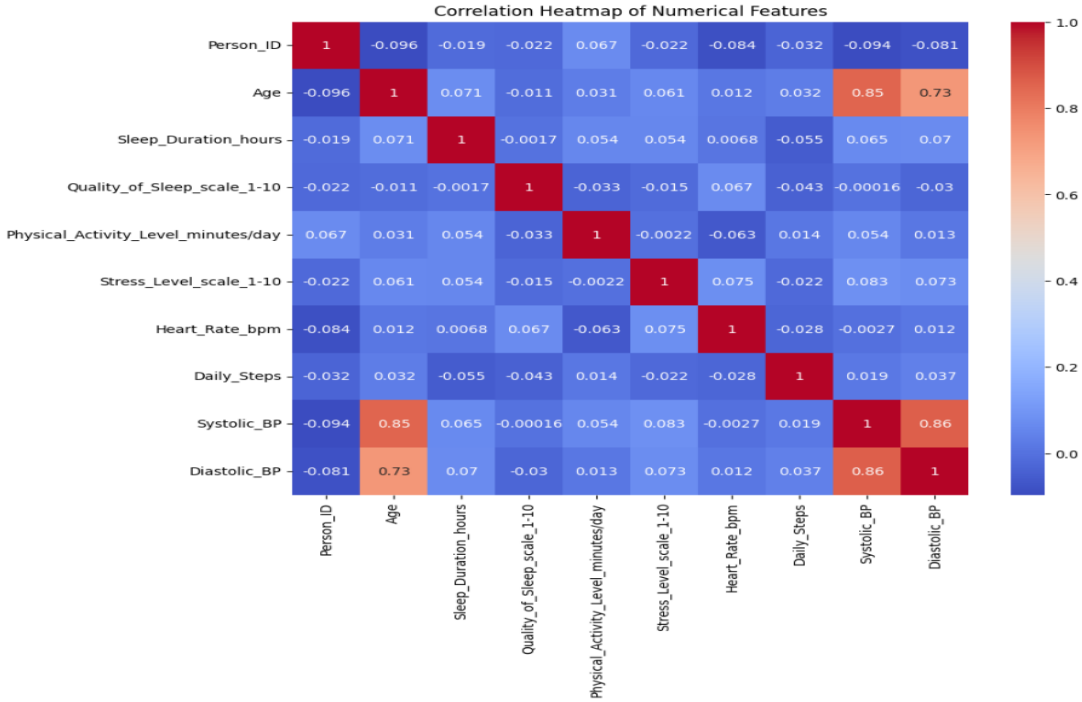

Figure 6. Correlation heatmap

3.2 Confusion Matrix and Class-wise Analysis

To evaluate the model's class-wise predictive performance, confusion matrices were analyzed for all three classifiers. Across models, the “None” class (no sleep disorder) was identified with the highest accuracy. However, performance declined for the “Insomnia” and “Sleep Apnea” classes, which were more frequently misclassified, particularly as each other. This misclassification suggests overlapping symptomatology and potentially shared underlying risk factors between these two conditions.

In multi-class classification, such misclassification patterns are common when

feature distributions are similar across minority classes, or when feature correlations are weakly discriminative. To explore this further, a correlation heatmap of numerical features was generated (Figure 6). The heatmap reveals relatively low correlation coefficients among key predictive features such as sleep duration, quality of sleep, stress level, and heart rate. For instance, the correlation between sleep duration and sleep quality is nearly zero (r = −0.0017), indicating that these two features, while intuitively related, behave independently in this dataset. Similarly, stress level shows weak positive correlations with both heart rate (r = 0.075) and sleep quality (r = 0.067), which may reduce the classifiers' ability to distinguish between insomnia and sleep apnea if they present similarly in terms of stress and fatigue.

Notably, age showed strong positive correlations with both systolic blood pressure (r = 0.85) and diastolic blood pressure (r = 0.73), suggesting that age-related physiological factors could play a stronger role in identifying sleep apnea, which often co-occurs with cardiovascular dysregulation. However, because these features are shared across multiple classes, their discriminative power remains limited without more granular clinical data.

From a machine learning standpoint, these findings explain the drop in recall and precision for the "Sleep Apnea" class, particularly in the Support Vector Machine (SVM) model, which lacks inherent mechanisms to handle feature redundancy or non-linear interactions effectively. Conversely, XGBoost and Random Forest, which use tree-based structures, are better equipped to manage weakly correlated features and still deliver relatively higher F1-scores for minority classes.

This reinforces the need for more sophisticated data modeling strategies to improve class separability and reduce misclassification errors. In particular, future studies could benefit from the application of advanced feature engineering techniques, such as introducing interaction terms or transforming variables based on domain knowledge, to reveal latent patterns that linear correlations may not capture. Furthermore, the inclusion of more targeted clinical features, for instance, oxygen saturation levels, snoring frequency, or apnea-hypopnea index that improve the discriminatory power between insomnia and sleep apnea, which often share overlapping symptom domains. Additionally, employing sampling techniques like the Synthetic Minority Over-sampling Technique (SMOTE) could help mitigate class imbalance by synthetically increasing the representation of minority classes during training, thus enhancing the model's sensitivity to underrepresented conditions.

3.3 Model Development and Performance Evaluation

To evaluate the predictive capabilities of the developed models, three supervised classification algorithms were implemented: Random Forest, Support Vector Machine (SVM), and XGBoost. Each model was trained on an 80% stratified training subset and evaluated on the remaining 20% test set using cross-validation and grid search for hyperparameter optimization.

Table 1 summarizes the performance of each model in terms of accuracy and macro-averaged F1-score, while Figure 10 provides a visual comparison. Among the models, XGBoost achieved the highest accuracy (0.90) and F1-score (0.88), followed by Random Forest (accuracy = 0.875, F1-score = 0.86). The SVM model recorded the lowest performance, with an accuracy of 0.825 and F1-score of 0.71. This trend indicates the superior generalization and adaptability of tree-based ensemble models especially gradient boosting in handling class imbalance and capturing complex feature interactions.

Table 1. Performance evaluation of the models

Model	Accuracy	F1-Score
Random Forest	0.875	0.86
XGBoost	0.900	0.88
SVM	0.825	0.71

The confusion matrices in Figure 7-Figure 9 further elaborate on model performance at the class level. All three models showed strong predictive capability for the "None" class (no sleep disorder), with XGBoost correctly classifying 55 instances, Random Forest 57, and SVM 58. However, classification performance dropped significantly for the "Insomnia" and "Sleep Apnea" classes, the minority classes in the dataset. Random Forest and SVM both failed to classify any cases of insomnia correctly, while XGBoost misclassified 14 of 16 insomnia cases as "None," only correctly identifying two. For sleep apnea, all three models classified six instances

correctly, but XGBoost again showed marginally better stability in misclassification patterns.

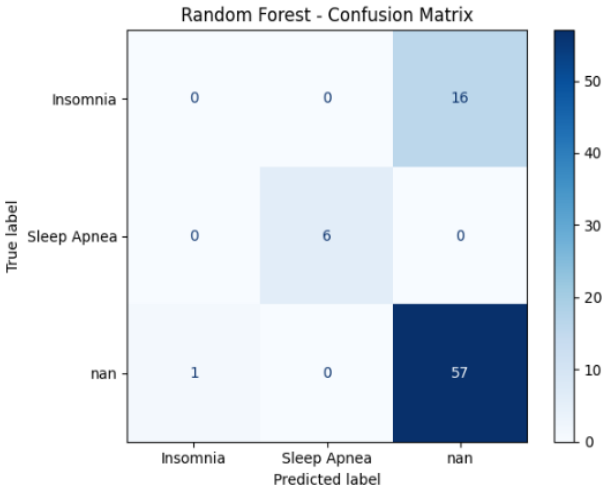

Figure 7. Random forest confusion matrix

The confusion matrix for XGBoost (Figure 8) reveals its slightly higher sensitivity in distinguishing insomnia from the "None" class compared to other models, though some confusion remains. This may be attributed to XGBoost's ability to reduce bias through sequential learning and capture non-linear boundaries. Conversely, SVM (Figure 9) struggled with minority class detection, failing to correctly identify any insomnia cases, likely due to its limitations in handling imbalanced, non-linearly separable data without kernel tuning or cost-sensitive weighting.

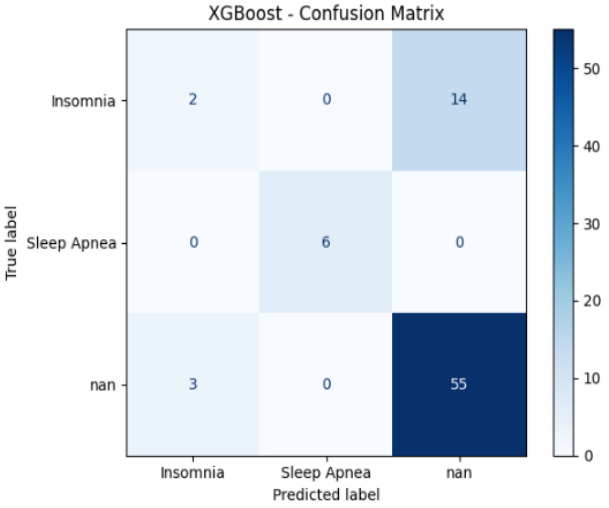

Figure 8. XGBoost confusion matrix

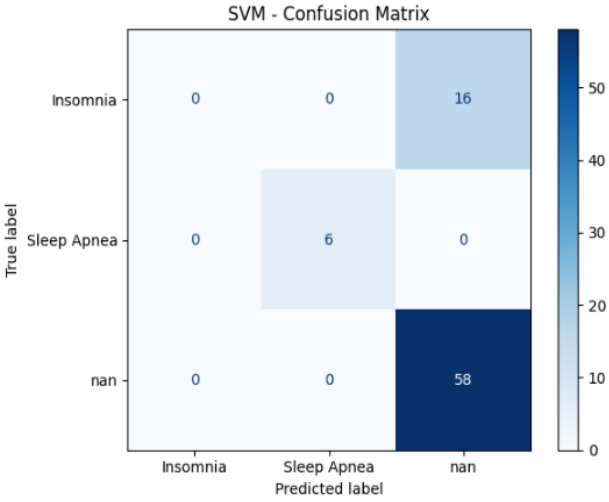

Figure 9. SVM confusion matrix

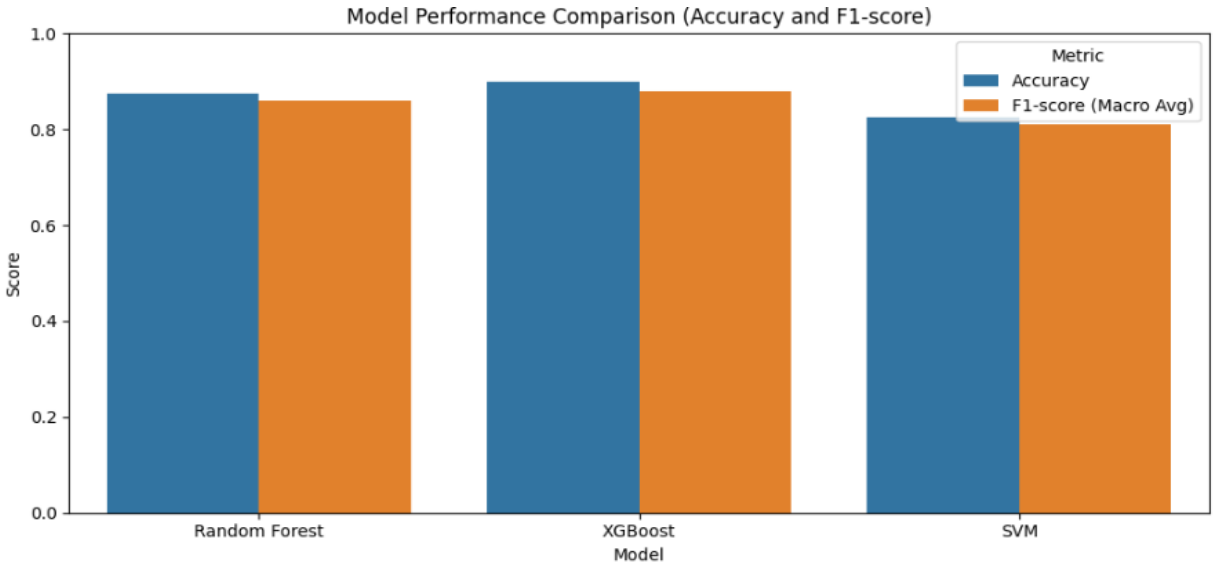

Figure 10. Model performance comparison

To evaluate overall classification performance, Table 1 summarizes the accuracy and macro-averaged F1-scores of all three models. These results are also visualized in Figure 10, which clearly highlights XGBoost as the top-performing model, followed by Random Forest, while SVM lags behind on both metrics. The close alignment between accuracy and F1-score in XGBoost and Random Forest indicates consistent performance across classes, while the lower F1-score in SVM reflects poor sensitivity toward minority classes, particularly insomnia.

4. Conclusion

This study demonstrates the feasibility and effectiveness of applying machine learning algorithms to predict sleep disorders based on a combination of occupational, lifestyle, and physiological data. Among the models tested, XGBoost achieved the highest accuracy (90%) and macro F1-score (88%), outperforming Random Forest and SVM in classifying sleep disorders. While all models exhibited strong performance in identifying individuals without sleep disorders, challenges persisted in correctly classifying minority classes such as insomnia and sleep apnea due to class imbalance and feature overlap. Correlation analysis confirmed that key predictors, such as sleep duration, stress level, and heart rate, had limited discriminative power across classes, necessitating more sophisticated modeling and richer datasets in future work. Additionally, occupational analysis revealed that manual laborers were more vulnerable to poor sleep outcomes, marked by elevated stress and shorter sleep duration. These findings underscore the value of integrating occupational health indicators into predictive pipelines for personalized and preventive sleep health interventions.

Acknowledgement

The authors gratefully acknowledge the support from the Department of Chemical Engineering Politeknik Negeri Bandung, and the Department of Engineering Physics and Nuclear Engineering, Universitas Gadjah Mada, for the collaboration to complete this research. Special thanks are also extended to the Kaggle community for openly sharing the Sleep Health and Lifestyle dataset that enabled this study.

References

[1] Lyons S, Strazdins L, Doan T. Work intensity and workers' sleep: A case of working Australians. Humanit Soc Sci Commun 2022;9:381.

[2] Khoshakhlagh, A. H., Sulaie, S. A., Cousins, R., Yazdanirad, S., & Laal, F., 2024. *Understanding the effect of occupational stress on sleep quality in firefighters: The modulating role of depression and burnout.* International Archives of Occupational and Environmental Health, Vol. 97, No. 9, pp. 1007-1016.

[3] Irawan, H.& Hartono, D., 2022. *Impact of urbanization on energy intensity in Indonesia: Spatial analysis.* Jurnal Perencanaan Pembangunan: The Indonesian Journal of Development Planning, Vol. 6, No. 2, pp. 202-215.

[4] Sathvik, S., Alsharef, A., Singh, A. K., Shah, M. A., & ShivaKumar, G., 2024. *Enhancing construction safety: Predicting worker sleep deprivation using machine learning algorithms.* Scientific Reports, Vol. 14, No. 1, p. 15716.

[5] Darvishi, E., Osmani, H., Aghaei, A., & Moloud, E. A., 2024. *Hidden risk factors and the mediating role of sleep in work-related musculoskeletal discomforts.* BMC Musculoskeletal Disorders, Vol. 25, No. 1, p. 256.

[6] Ciharova, M. *et al.*, 2025. *Machine-learning detection of stress severity expressed on a continuous scale using acoustic, verbal, visual, and physiological data: Lessons learned.* Frontiers in Psychiatry, Vol. 16, p. 1548287.

[7] Ha, S. *et al.*, 2023. *Predicting the risk of sleep disorders using a machine learning–based simple questionnaire: Development and validation study.* Journal of medical Internet research, Vol. 25, p. e46520.

[8] Mostafa Monowar, M. *et al.*, 2025. *Advanced sleep disorder detection using multi-layered ensemble learning and advanced data balancing techniques.* Frontiers in Artificial Intelligence, Vol. 7, p. 1506770.

[9] Khanmohmmadi, S., Khatibi, T., Tajeddin, G., Akhondzadeh, E., & Shojaee, A., 2025. *Revolutionizing sleep disorder diagnosis: A multi-task learning approach optimized with genetic and q-learning techniques.* Scientific Reports, Vol. 15, No. 1, p. 16603.

[10] Sleep Heath Life Style n.d. https://www.kaggle.com/datasets/dawoodhuss227/sleep-heath-life-style (accessed July 30, 2025).

[11] Alqethami, N. A., 2025. *Machine learning approaches for sleep disorder classification: Insights from optuna-based hyperparameter tuning.* Journal of Medical Artificial Intelligence.

[12] Markov, K., Elgendi, M., Birrer, V., & Menon, C., 2025. *Interpretable feature-based machine learning for automatic sleep detection using photoplethysmography.* npj Biosensing, Vol. 2, No. 1, p. 24.

[13] Wara, T. U., Fahad, A. H., Das, A. S., & Shawon, M. M. H., 2025. *A systematic review on sleep stage classification and sleep disorder detection using artificial intelligence.* Heliyon, Vol. 11, No. 12.

[14] Mostafa Monowar, M. *et al.*, 2025. *Advanced sleep disorder detection using multi-layered ensemble learning and advanced data balancing techniques.* Frontiers in Artificial Intelligence, Vol. 7, p. 1506770.

[15] Zovko, K. *et al.*, 2025. *Advanced data framework for sleep medicine applications: Machine learning-based detection of sleep apnea events.* Applied Sciences, Vol. 15, No. 1, p. 376.

[16] Dritsas, E.& Trigka, M., 2024. *Utilizing multi-class classification methods for automated sleep disorder prediction.* Information, Vol. 15, No. 8, p. 426.