Journal of Electrical Engineering and Computer (JEECOM)
Vol. 8, No. 2 (2026), DOI: 10.33650/jeecom.v8i2.17445
p-ISSN: 2715-0410; e-ISSN: 2715-6427

Predicting Sleep Quality Using Hybrid Stacking Ensemble Of Random Forest And Catboost With Shap Interpretability

Baitullah Hakim[1*], Muhammad Faisal[2], Mochamad Imamudin[3]
[1,2,3] Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia

Article Info

Article history:

Received September 30, 2026
Revised October 05, 2026
Accepted October 06, 2026

Keywords:

sleep quality prediction, Random Forest, CatBoost, stacking ensemble, Optuna, SHAP, machine learning

ABSTRACT

Sleep quality is an important indicator of physical and mental well-being, yet its prediction from lifestyle variables remains challenging because behavioral and physiological factors interact nonlinearly. This study develops a hybrid stacking ensemble that combines Random Forest and CatBoost as base learners and Logistic Regression as the meta-classifier for three-class sleep-quality prediction (Poor, Moderate, Good). The Sleep Health and Lifestyle Dataset contains 374 records, of which only 109 are unique, and is preprocessed through blood-pressure decomposition, derived-feature construction, categorical encoding, and numerical standardization. Models are evaluated with Accuracy, Precision, Recall, F1-Score, and multiclass ROC-AUC under Stratified 10-Fold Cross Validation, both in its standard form and in a duplicate-aware form in which identical records are kept in the same fold. Under the duplicate-aware protocol, Random Forest achieved 97.86% Accuracy, 91.06% macro F1-Score, and 0.9970 ROC-AUC; CatBoost achieved 97.33%, 90.70%, and 0.9930; and Hybrid Stacking achieved 97.06%, 90.51%, and 0.9784. Recall for the minority Poor class (n = 12) was 66.7% for all three models, and stacking did not improve on its base learners. SHAP analysis identifies Stress Level and Sleep Duration as the two dominant features. The results are specific to this small, largely duplicated dataset and do not support clinical use.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Baitullah Hakim
Magister Informatika, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia
Email: 250605210025@student.uin-malang.ac.id ; Orcid Id: https://orcid.org/0009-0000-0175-251X

1. INTRODUCTION

Sleep is a fundamental physiological process associated with cognitive function, emotional regulation, metabolic health, cardiovascular function, and immune regulation. The American Academy of Sleep Medicine and Sleep Research Society recommend that healthy adults obtain at least seven hours of sleep per night on a regular basis [1]. Beyond duration, sleep quality includes the continuity, efficiency, ease of sleep initiation, and restorative effect of sleep [2]. Consequently, sleep quality is a multidimensional outcome that is influenced by behavioral, physiological, and environmental factors.

Artificial intelligence (AI) and machine learning (ML) provide an opportunity to model interactions among lifestyle and health variables without requiring manually specified decision rules. In sleep medicine, ML has been applied to physiological signals, questionnaires, wearable-device data, and clinical variables. Sathish et al. used convolutional neural networks to predict sleep quality from sleep-related data [3], while Ha et al. developed a machine-learning-based questionnaire approach for predicting sleep-disorder risk [4]. Han and Oh compared several ML algorithms for obstructive sleep apnea severity prediction and demonstrated the usefulness of ensemble methods for heterogeneous clinical data [5]. Clinical perspectives and comparative studies further discuss the potential and the limitations of ML in sleep medicine [6], [7], and boosting algorithms have been compared for sleep-disorder prediction [8].

ISSN: 2715-6427

Recent work has also highlighted the potential of tree-based ensemble learning. Zaman et al. proposed SleepBoost, a multi-level ensemble combining CatBoost, LightGBM, and Random Forest for sleep-stage classification [9]. Sakal et al. used wearable-device data and CatBoost for proactive sleep-efficiency forecasting [10]. Kaur and Neeru reported high classification performance from ensemble models using clinical and behavioral sleep-related variables [11]. These studies demonstrate the relevance of ensemble learning, but their objectives, data sources, and target definitions differ from personal sleep-quality classification using lifestyle variables.

Random Forest and CatBoost offer complementary characteristics for tabular data. Random Forest aggregates randomized decision trees through bagging and is known for robust generalization [12]. CatBoost applies gradient boosting with ordered boosting, a mechanism designed to reduce prediction shift and overfitting, and it supports categorical variables natively [13]. The predictors in the dataset used here are predominantly numerical (six numerical versus four categorical variables), and the categorical variables are label-encoded in the present pipeline; CatBoost is therefore not selected for categorical handling alone, but because ordered boosting and symmetric (oblivious) trees act as strong regularizers on small tabular datasets, and tree-based models remain highly competitive on tabular data in general [14]. Combining heterogeneous learners through stacking can allow a meta-classifier to learn how to combine their predictions [15]. However, increased architectural complexity does not guarantee improved generalization, especially for a small dataset.

Another important issue in health-oriented ML is interpretability. A model can produce predictions without clearly showing which variables contributed to those predictions. SHapley Additive exPlanations (SHAP) provides a game-theoretic framework for attributing a model output to individual features [16], and its TreeExplainer variant enables efficient explanations for tree-based models [17]. In a health-related setting, such explanations can provide additional context for model behavior, although feature attribution must not be interpreted as evidence of causality.

Based on these considerations, this study investigates a Hybrid Ensemble Learning framework that combines Random Forest and CatBoost through a Stacking Ensemble with Logistic Regression as the meta-classifier. Optuna is used for hyperparameter optimization, and SHAP is applied to the tree-based base learners. The specific objectives are to: (1) construct a three-class sleep-quality prediction model from lifestyle and health variables; (2) compare the performance of Random Forest, CatBoost, and the Hybrid Ensemble; (3) obtain optimized hyperparameter configurations using Optuna; and (4) identify prominent features in the model-based SHAP explanations. The contribution of this study is therefore methodological and empirical: it reports both the capabilities and the limitations of the proposed hybrid architecture on a small public tabular dataset.

2. METHOD

2.1 Dataset and Variables

The study uses the Sleep Health and Lifestyle Dataset, a publicly available secondary dataset obtained from Kaggle [18]. Unlike large-scale Kaggle collections, this is a compact, single-source tabular dataset: the original file contains 374 records and 13 columns, and 11 columns are retained after excluding the identifier (Person ID) and Sleep Disorder. The small number of records is therefore a characteristic of the dataset itself and is discussed as a limitation in Section 3.8. The target is Quality of Sleep, and the predictors are Gender, Age, Occupation, Sleep Duration, Physical Activity Level, Stress Level, BMI Category, Blood Pressure, Heart Rate, and Daily Steps, as summarized in Table 1. The original Quality of Sleep score ranges from 4 to 9 and is transformed into three classes: Poor (4-5), Moderate (6-7), and Good (8-9).

Table 1. Variables used in the sleep-quality classification task.

Feature group	Variables	Role
Categorical	Gender, Occupation, BMI Category, Age_Group	Predictors
Numerical	Age, Sleep Duration, Physical Activity Level, Stress Level, Heart Rate, Daily Steps, Systolic_BP, Diastolic_BP, Sleep_Activity_Ratio	Predictors
Target	Quality of Sleep	3-class target

2.2 Data Preprocessing

The preprocessing pipeline consists of the following steps. Blood Pressure is decomposed into Systolic_BP and Diastolic_BP. Two derived variables are constructed as part of feature engineering: Age_Group, which categorizes Age as four ordinal groups with left-closed intervals, namely below 30 years (label 0), 30-44 years (label 1), 45-59 years (label 2), and 60 years or older (label 3), and Sleep_Activity_Ratio, defined as Sleep Duration (hours) divided by Physical Activity Level (minutes/day), computed as Sleep Duration / (Physical Activity Level + 10^{-6}), where the small constant prevents division by zero. Missing numerical values are imputed using the median and missing categorical values using the mode. Categorical variables are transformed

to numerical representations using LabelEncoder, while numerical variables are standardized using StandardScaler. No additional feature-selection step is applied in the final reproducible pipeline; all retained predictors are processed consistently before model training.

Preprocessing steps that learn parameters from the data (standardization and the inner folds of the stacking classifier) are fitted within each training fold to avoid information leakage. The predictors contain no missing values, so imputation had no effect. The target distribution consists of 12 Poor records (3.21%), 182 Moderate records (48.66%), and 180 Good records (48.13%). Because the classes are imbalanced, Stratified 10-Fold Cross Validation is used to preserve class proportions across folds.

Inspection of the dataset showed that only 109 of the 374 records are unique when all predictors and the target are compared; the remaining records are exact duplicates. If duplicates fall into both the training and the test folds, performance can be overestimated. Two protocols are therefore reported: (i) standard Stratified 10-Fold Cross Validation (shuffled, seed 42), and (ii) duplicate-aware Stratified Group 10-Fold Cross Validation, in which identical records form one group and are never split between training and test folds. Protocol (ii) is treated as the primary evaluation.

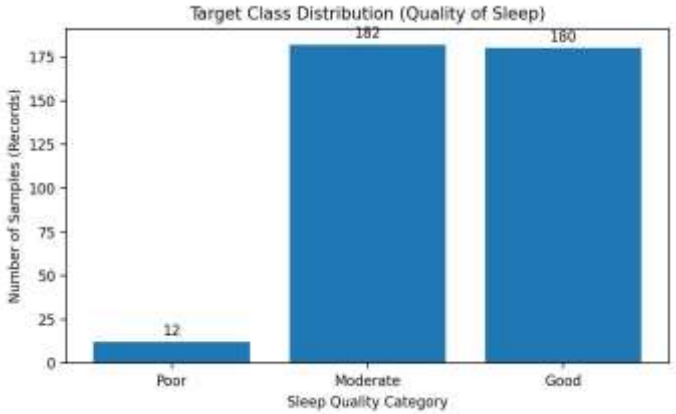


Figure 1. Distribution of the three Quality of Sleep classes.

2.3 Hybrid Stacking Ensemble

The proposed architecture contains two base learners and one meta-classifier. Random Forest and CatBoost independently learn the mapping from the processed features to the three sleep-quality classes. Their predicted class probabilities are then used as meta-features for Logistic Regression. The meta-classifier produces the final class prediction.

For each validation fold, the base learners are fitted on the training data, and their out-of-fold class probabilities, generated by an inner 5-fold cross-validation, are passed to the meta-classifier. The process is repeated for every outer fold.

2.4 Hyperparameter Optimization with Optuna

Optuna is used to search for promising hyperparameter configurations for Random Forest and CatBoost. The implementation performs 10 trials for Random Forest and 10 trials for CatBoost. The optimization objective is the macro F1-score under Stratified 10-Fold Cross Validation, and the selected configurations are then used in the subsequent model experiments. Optuna was selected because it provides automated search over a defined parameter space and was introduced as a next-generation hyperparameter optimization framework [19]. The best configurations obtained are listed in Table 2. Tuning was performed once and the resulting configurations were reused, without re-tuning, in the evaluation reported here; the performance estimates are therefore not nested and may be slightly optimistic.

Table 2. Hyperparameter configurations obtained from Optuna.

Model	Hyperparameter	Best value
Random Forest	n_estimators	199
Random Forest	max_depth	11
Random Forest	min_samples_split	6
Random Forest	min_samples_leaf	1
CatBoost	iterations	357
CatBoost	learning_rate	0.0281
CatBoost	depth	5
CatBoost	l2_leaf_reg	0.00615

ISSN: 2715-6427

2.5 Model Evaluation

Performance is evaluated using Accuracy, Precision, Recall, F1-Score, and multiclass ROC-AUC [21]–[23]. Accuracy measures the proportion of correctly classified observations. Precision reflects the correctness of positive predictions, Recall reflects the ability to identify observations belonging to a class, and F1-Score summarizes Precision and Recall through their harmonic mean [22]. ROC-AUC is used as an overall discrimination measure for the multiclass setting [23]. The primary evaluation uses duplicate-aware Stratified 10-Fold Cross Validation (Section 2.2) [20]. Because the largest class (Moderate) contains 182 of the 374 records (48.66%), this majority-class proportion, together with the chance level of approximately 33.3% for three classes, is used as a reference when interpreting Accuracy. Precision, Recall, and F1-Score are macro-averaged, and per-class recall is reported because the Poor class contains only 12 records.

2.6 SHAP based Interpretation

SHAP is applied using TreeExplainer [17] to the CatBoost and Random Forest models. The analysis uses SHAP values to quantify the contribution of features to model outputs. The mean absolute SHAP magnitude provides a global indication of feature importance, while the sign and distribution of individual SHAP values can be used to examine directional contributions. Importantly, SHAP explains model behavior; it does not establish that a feature causes a change in sleep quality. SHAP values are computed on models fitted to the full standardized dataset, so they describe the fitted models rather than out-of-sample behavior; features are ranked by mean absolute SHAP value averaged over the three classes.

3. RESULTS AND DISCUSSION

3.1 Dataset Preparation and Class Distribution

The dataset contains 374 records. The preprocessing stage separated blood pressure into two numerical features, generated Age_Group and Sleep_Activity_Ratio, handled missing values through the defined preprocessing pipeline, encoded categorical variables, and standardized numerical variables. The target classes were derived from Quality of Sleep: Poor (scores 4-5), Moderate (6-7), and Good (8-9).

Table 3. Distribution of the Quality of Sleep classes.

Class	Records	Proportion
Poor	12	3.21%
Moderate	182	48.66%
Good	180	48.13%

The class distribution contains 12 Poor records (3.21%), 182 Moderate records (48.66%), and 180 Good records (48.13%). The largest class therefore represents 48.66% of the dataset. Stratified validation is used to preserve the class proportions during model evaluation. Only 109 of the 374 records are unique (Section 2.2).

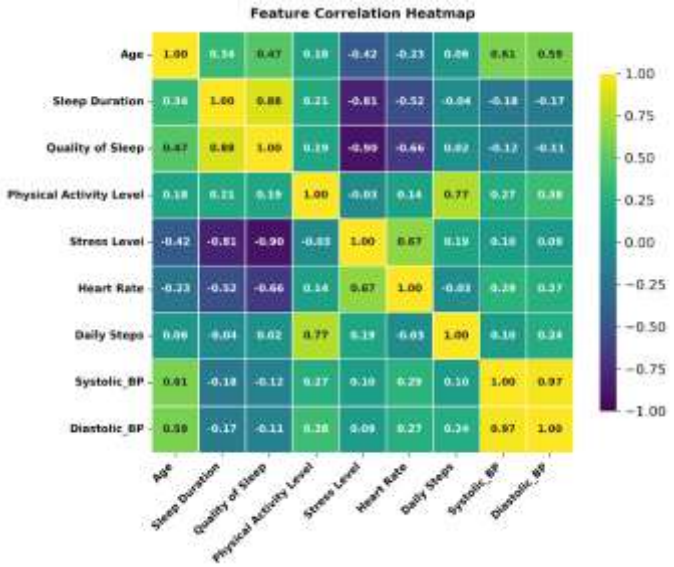


Figure 2. Correlation matrix of the preprocessed numerical features.

The correlation matrix in Figure 2 is used as an exploratory description of the numerical feature structure. It should not be interpreted as evidence of causal relationships. The visualization indicates that the selected

numerical variables do not form a single strongly correlated linear block, supporting the use of nonlinear tree-based learners for further experimentation.

3.2 Optimized Model Configurations

As shown in Table 2, the final documented configurations are 199 trees, maximum depth 11, minimum samples split 6, and minimum samples leaf 1 for Random Forest. For CatBoost, the configuration contains 357 iterations, a learning rate of approximately 0.0281, depth 5, and the documented L2 regularization value. These configurations are evaluated under the same Stratified 10-Fold Cross Validation protocol.

3.3. Hybrid Stacking Evaluation

Table 4 summarizes the performance of the Hybrid Stacking model under the primary, duplicate-aware Stratified 10-Fold Cross Validation protocol.

Table 4. Performance of the Hybrid Stacking model under the primary (duplicate-aware) evaluation protocol.

Metric	Hybrid Stacking
Accuracy	0.9706 (97.06%)
Precision (macro)	0.9448 (94.48%)
Recall (macro)	0.8760 (87.60%)
F1-Score (macro)	0.9051 (90.51%)
ROC-AUC	0.9784

Under the duplicate-aware protocol, the Hybrid Stacking model achieved an Accuracy of 97.06%, macro Precision of 94.48%, macro Recall of 87.60%, macro F1-Score of 90.51%, and ROC-AUC of 0.9784 (Table 4). The gap between Accuracy and macro Recall/F1-Score arises from the minority Poor class: only 8 of its 12 records were identified correctly (Table 7). Because the two large classes make up 96.8% of the records, Accuracy alone overstates performance on the minority class. The high Accuracy is plausible because Quality of Sleep is strongly associated with Sleep Duration (r = 0.88) and Stress Level (r = -0.90) in this dataset.

3.4 Comparison with Base Models

Table 5 compares Random Forest, CatBoost, and the Hybrid Stacking model under the duplicate-aware protocol, and Figure 3 visualizes the same metrics. Table 6 reports the same models under standard Stratified 10-Fold Cross Validation, in which duplicate records may appear in both training and test folds, and Table 7 reports class-specific results for the primary protocol.

Table 5. Performance comparison of Random Forest, CatBoost, and Hybrid Stacking under the duplicate-aware protocol (macro-averaged Precision, Recall, and F1).

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Random Forest	0.9786	0.9506	0.8815	0.9106	0.9970
CatBoost	0.9733	0.9469	0.8778	0.9070	0.9930
Hybrid Stacking	0.9706	0.9448	0.8760	0.9051	0.9784

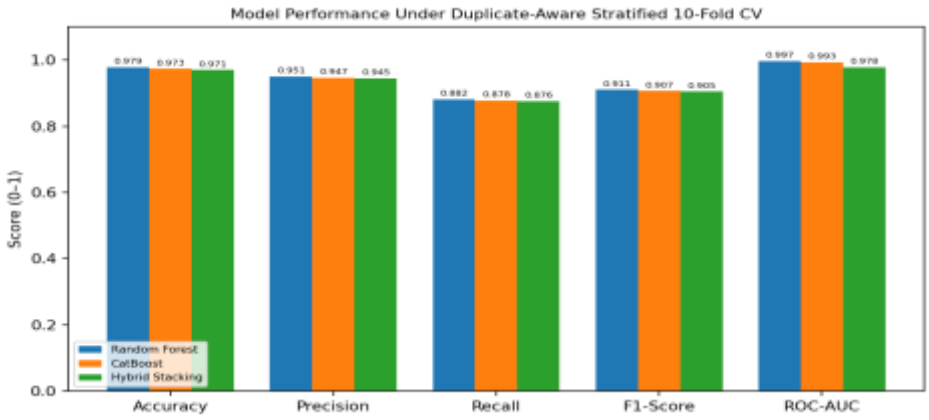


Figure 3. Model comparison under the duplicate-aware protocol.

Under the duplicate-aware protocol, Random Forest achieved the highest Accuracy (97.86%), macro F1-Score (91.06%), and ROC-AUC (0.9970), followed by CatBoost and Hybrid Stacking (Table 5). The differences among the three models are small (about 0.8 percentage points of Accuracy between the best and the worst)

ISSN: 2715-6427

and rest on only 109 unique observations and 12 Poor records, so they should not be read as a reliable ranking. In particular, stacking two learners did not improve on its base learners, which is consistent with the principle that ensemble complexity does not guarantee better generalization on small datasets [15].

Table 6 shows that standard Stratified 10-Fold Cross Validation gives higher values (Accuracy of 99.20% for all three models and macro F1-Score between 96.86% and 98.21%), because duplicate records can appear in both training and test folds. Macro F1-Score decreases by roughly 6 to 8 percentage points when duplicates are grouped, and Poor-class recall decreases from 91.7-100% to 66.7% (Table 7). The duplicate-aware results are therefore considered more representative of performance on unseen individuals, although they remain optimistic because hyperparameters were tuned on the same data.

Table 6. Performance under standard Stratified 10-Fold Cross Validation (duplicates may appear in both training and test folds).

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Random Forest	0.9920	0.9686	0.9685	0.9686	0.9988
CatBoost	0.9920	0.9707	0.9945	0.9821	0.9992
Hybrid Stacking	0.9920	0.9707	0.9945	0.9821	0.9989

Table 7. Class-specific recall under the duplicate-aware protocol.

Model	Recall (Poor)	Recall (Moderate)	Recall (Good)	Poor found (of 12)	F1 (macro)
Random Forest	0.667	0.995	0.983	8	0.9106
CatBoost	0.667	0.989	0.978	8	0.9070
Hybrid Stacking	0.667	0.978	0.983	8	0.9051

For each model, the four Poor records that were missed were all assigned to the Moderate class and none to the Good class, so adjacent classes are confused but distant ones are not. Because the Poor class holds only 12 records, each record corresponds to 8.3 percentage points of recall, and the estimate is highly uncertain.

3.5 SHAP Interpretation

Figures 4 and 5 present the SHAP feature-importance summaries of the CatBoost and Random Forest models, respectively.

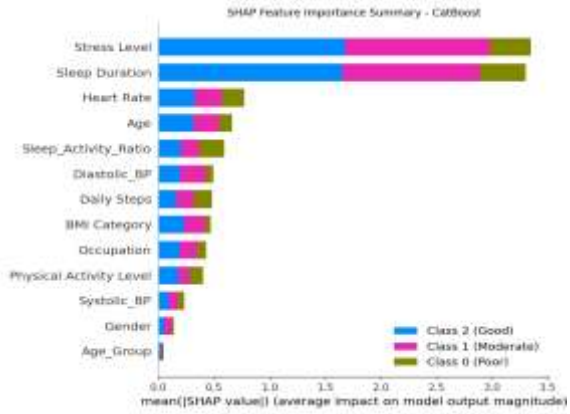


Figure 4. SHAP feature-importance summary for CatBoost.

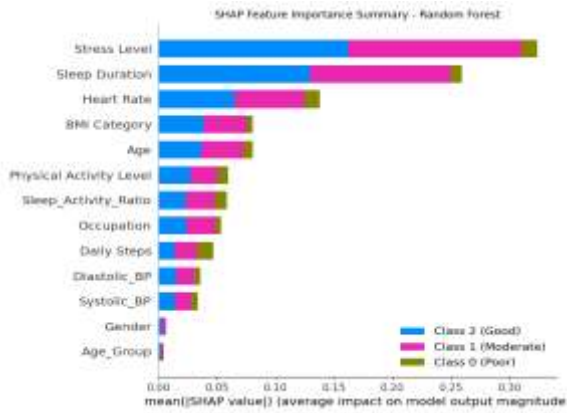


Figure 5. SHAP feature-importance summary for Random Forest.

The SHAP summaries in Figures 4 and 5 show Stress Level and Sleep Duration as the two dominant features in both models. In CatBoost, their mean absolute SHAP values (1.12 and 1.10, averaged over the classes) are more than four times that of the next feature, Heart Rate (0.26); in Random Forest, Stress Level (0.108) and Sleep Duration (0.086) lead Heart Rate (0.046) by roughly a factor of two. Age and several secondary or derived features follow with much smaller contributions, and Daily Steps contributes little (its correlation with Quality of Sleep is 0.02). Because Sleep_Activity_Ratio is derived from Sleep Duration, its contribution partly overlaps with that of Sleep Duration. The prominence of Stress Level is in line with the established association between psychosocial stress and impaired sleep [24].

The interpretation should remain model-specific. A high SHAP contribution means that the feature contributes substantially to the model output under the evaluated data distribution; it does not mean that changing the feature will necessarily cause a corresponding change in sleep quality. This distinction is particularly important in health-related applications.

3.6 Exploratory Relationship between Sleep Duration and Stress

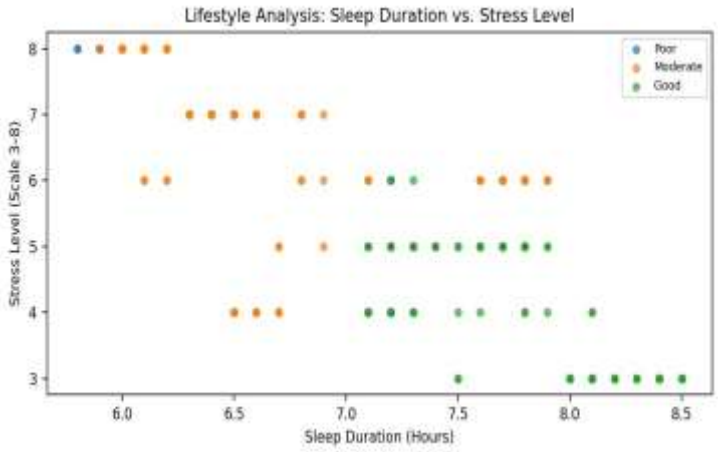


Figure 6. Sleep Duration and Stress Level by sleep-quality class.

The scatter plot in Figure 6 shows that the three target categories occur across different combinations of Sleep Duration and Stress Level. This supports the view that classification cannot be adequately explained by a single variable pair and motivates the use of multivariate models. Nevertheless, the visualization is descriptive and does not establish causality.

3.7 Comparison with Previous Studies

The performance obtained in this study should be compared with related studies cautiously because the datasets, target definitions, feature sets, and validation protocols differ. Published studies have reported different performance levels under their own experimental settings; therefore, the values in this study are interpreted only within the evaluation protocol used here.

3.8 Implications and Limitations

From a methodological perspective, the study demonstrates a workflow from preprocessing and hyperparameter configuration to ensemble learning and explainability. From an application perspective, the model is a prototype for lifestyle-based sleep-quality classification and is not a substitute for clinical assessment.

A key limitation is the small sample size: the dataset contains 374 records, but only 109 are unique, and the Poor class has only 12 records. The dataset is a single public secondary source with only 109 unique records; the absence of an independently documented collection protocol limits external validity, so the results should not be generalized to real populations or used clinically. In addition, hyperparameters were tuned by cross-validation on the same data used for performance estimation, so the estimates may be optimistically biased; nested cross-validation is recommended for model selection [26].

Future work should evaluate the pipeline on larger, independently collected sleep datasets with a documented collection protocol, use nested cross-validation to separate tuning from evaluation, and report fold-level metrics, confusion matrices, and calibration. The Poor class should be examined with more samples or cost-

sensitive methods, and further work can compare additional ensemble architectures and examine whether the SHAP rankings remain stable across independent datasets.

4. CONCLUSION

This study implemented a Hybrid Ensemble Learning framework combining Random Forest and CatBoost with Logistic Regression as a stacking meta-classifier for three-class sleep-quality prediction. The pipeline used the 374-record Sleep Health and Lifestyle Dataset (109 unique records), preprocessing through blood-pressure decomposition, feature engineering, categorical encoding, and numerical standardization, and Stratified 10-Fold Cross Validation in standard and duplicate-aware forms. SHAP was used to describe feature contributions in the tree-based learners.

Under the duplicate-aware protocol, Random Forest achieved 97.86% Accuracy and 91.06% macro F1-Score, CatBoost 97.33% and 90.70%, and Hybrid Stacking 97.06% and 90.51% (ROC-AUC of 0.9970, 0.9930, and 0.9784). Under standard Stratified 10-Fold Cross Validation all three models reached 99.20% Accuracy, which is inflated by duplicate records. Recall for the minority Poor class was 66.7% for all models, and stacking did not improve on its base learners. These figures describe the observed results on this dataset and should not be interpreted as evidence of clinical validity or generalization beyond the evaluated data.

SHAP-based explanations identify Stress Level and Sleep Duration as the dominant features in both the CatBoost and Random Forest models, followed by Heart Rate and other features with much smaller contributions. These results describe model-specific feature contributions and do not establish causal relationships.

ACKNOWLEDGMENTS

The authors would like to thank the academic environment that supported the development of this research and the open-data and open-source communities whose resources enabled the experiment. The authors also acknowledge the use of the Sleep Health and Lifestyle Dataset and the software libraries used for machine learning, optimization, visualization, and explainability.

REFERENCES

[1] K. Ramar et al., “Sleep is essential to health: An American Academy of Sleep Medicine position statement,” Journal of Clinical Sleep Medicine, vol. 17, no. 10, pp. 2115–2119, 2021, doi: 10.5664/jcsm.9476.

[2] M. Fabbri, A. Beracci, M. Martoni, D. Meneo, L. Tonetti, and V. Natale, “Measuring subjective sleep quality: A review,” International Journal of Environmental Research and Public Health, vol. 18, no. 3, 1082, 2021, doi: 10.3390/ijerph18031082.

[3] V. R. K. Sathish, W. L. Woo, and E. S. L. Ho, “Predicting sleeping quality using convolutional neural networks,” arXiv:2204.13584, 2022, doi: 10.48550/arXiv.2204.13584.

[4] S. Ha et al., “Predicting the risk of sleep disorders using a machine learning–based simple questionnaire: Development and validation study,” Journal of Medical Internet Research, vol. 25, e46520, 2023, doi: 10.2196/46520.

[5] H. Han and J. Oh, “Application of various machine learning techniques to predict obstructive sleep apnea syndrome severity,” Scientific Reports, vol. 13, 6379, 2023, doi: 10.1038/s41598-023-33170-7.

[6] A. Bandyopadhyay and C. Goldstein, “Clinical applications of artificial intelligence in sleep medicine: A sleep clinician’s perspective,” Sleep and Breathing, vol. 27, no. 1, pp. 39–55, 2023, doi: 10.1007/s11325-022-02592-4.

[7] R. Alazaidah, G. Samara, M. Aljaidi, M. Haj Qasem, A. Alsarhan, and M. Alshammari, “Potential of machine learning for predicting sleep disorders: A comprehensive analysis of regression and classification models,” Diagnostics, vol. 14, no. 1, p. 27, 2024, doi: 10.3390/diagnostics14010027.

[8] A. B. Mawardi, R. S. Pradini, and M. S. Haris, “Komparasi algoritma boosting untuk prediksi gangguan tidur,” Jurnal Informatika dan Teknik Elektro Terapan, vol. 13, no. 3, 2025, doi: 10.23960/jitet.v13i3.7281.

[9] A. Zaman, S. Kumar, S. Shatabda, I. Dehzangi, and A. Sharma, “SleepBoost: A multi-level tree-based ensemble model for automatic sleep stage classification,” Medical & Biological Engineering & Computing, vol. 62, no. 9, pp. 2769–2783, 2024, doi: 10.1007/s11517-024-03096-x.

[10] C. Sakal, T. Chen, W. Xu, W. Zhang, Y. Yang, and X. Li, “Towards proactively improving sleep: Machine learning and wearable device data forecast sleep efficiency 4–8 hours before sleep onset,” SLEEP, vol. 48, no. 8, zsaf113, 2025, doi: 10.1093/sleep/zsaf113.

[11] A. Kaur and N. Neeru, "Automated detection of sleep disorders using ensemble learning on clinical and behavioral features," Sleep Medicine Research, vol. 17, no. 1, pp. 6–20, 2026, doi: 10.17241/smr.2025.03083.
[12] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," Advances in Neural Information Processing Systems, vol. 35, 2022, doi: 10.48550/arXiv.2207.08815.
[13] N. Safaei et al., "E-CatBoost: An efficient machine learning framework for predicting ICU mortality using the eICU Collaborative Research Database," PLoS ONE, vol. 17, no. 5, e0262895, 2022, doi: 10.1371/journal.pone.0262895.
[14] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," in Advances in Neural Information Processing Systems, vol. 35, 2022, doi: 10.48550/arXiv.2207.08815.
[15] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," IEEE Access, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.
[16] H. Chen, S. M. Lundberg, and S.-I. Lee, "Explaining a series of models by propagating Shapley values," Nature Communications, vol. 13, 4512, 2022, doi: 10.1038/s41467-022-31384-3.
[17] R. Dwivedi et al., "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," ACM Computing Surveys, vol. 55, no. 9, Article 194, 2023, doi: 10.1145/3561048.
[18] L. Tharmalingam, "Sleep Health and Lifestyle Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset
[19] B. Bischl et al., "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," WIREs Data Mining and Knowledge Discovery, vol. 13, no. 2, e1484, 2023, doi: 10.1002/widm.1484.
[20] R. D. King, O. I. Orhobor, and C. C. Taylor, "Cross-validation is safe to use," Nature Machine Intelligence, vol. 3, p. 276, 2021, doi: 10.1038/s42256-021-00332-z.
[21] S. A. Hicks et al., "On evaluation metrics for medical applications of artificial intelligence," Scientific Reports, vol. 12, 5979, 2022, doi: 10.1038/s41598-022-09954-8.
[22] J. Opitz, "A closer look at classification evaluation metrics and a critical reflection of common evaluation practice," Transactions of the Association for Computational Linguistics, vol. 12, pp. 820–836, 2024, doi: 10.1162/tacl_a_00675.
[23] Z. Yang et al., "Learning with multiclass AUC: Theory and algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 11, pp. 7747–7763, 2022, doi: 10.1109/TPAMI.2021.3101125.
[24] M. Gardani et al., "A systematic review and meta-analysis of poor sleep, insomnia symptoms and stress in undergraduate students," Sleep Medicine Reviews, vol. 61, 101565, 2022, doi: 10.1016/j.smrv.2021.101565.
[25] S. Atoui et al., "Daily associations between sleep and physical activity: A systematic review and meta-analysis," Sleep Medicine Reviews, vol. 57, 101426, 2021, doi: 10.1016/j.smrv.2021.101426.
[26] J. Wainer and G. Cawley, "Nested cross-validation when selecting classifiers is overzealous for most practical applications," Expert Systems with Applications, vol. 182, 115222, 2021, doi: 10.1016/j.eswa.2021.115222.