iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80272

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Machine Learning-Based Analysis of Stress Using Sleep Behavior Data

Snehal K. Kulkarni[1], Prasad T. Goyal[2], Tejaswini J. Parabkar[3], Harsha N.Une[4], Vaishnavi J.Davari[5], Sangram A.Killedar[6]

[1, 5]*Assistant Professor, Department of Computer Science, SVLM Titave,*
[3, 4]*Assistant Professor, Department of Mass Media, SVLM Titave, Librarian, Shahid Virpatni Laxmi Mahavidyalaya Titave,*
[2]*Assistant Professor, Department of Computer Science, Shivaji University, Kolhapur*

Abstract: ***Sleep is essential for human health, greatly influencing mental performance, emotional stability, and general well-being. Lack of sleep, which pertains to inadequate rest, is often linked to heightened stress levels. Nevertheless, the connection between stress and different influencing elements is intricate, complicating efforts to evaluate and forecast accurately.***
This study examines the use of machine learning techniques to predict stress levels using sleep-related and behavioral variables. The dataset includes 60 participants and 14 variables such as sleep duration, sleep quality, cognitive performance, emotional regulation, and lifestyle factors. To tackle the issue of a restricted dataset, methods for creating synthetic data were employed to increase the sample size while maintaining statistical correlations.
Four machine learning models—Linear Regression, Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost)—were applied and tested using multiple performance metrics. The results show that although synthetic data improves model training, predicting stress using only sleep-related variables is still limited. Among the models, XGBoost performed relatively better but still showed modest predictive capability. The study shows the need for adding additional physiological and environmental factors for more better stress prediction.
Keywords: Sleep Deprivation, Stress Prediction, Machine Learning, Synthetic Data, XGBoost, Behavioral Data.

I. INTRODUCTION

Sleep has an important part in maintaining both physical and internal health. Indeed though it's important, numerous individualities witness not enough or poor- quality sleep because of ultramodern life factors similar as advanced workload, screen exposure, and irregular routines. One of the most direct goods of sleep privation is advanced stress, which can further impact productivity, emotional balance, and decision- timber.

Stress is naturally complex and can not be measured fluently through a single parameter. It's impact by cerebral, physiological, and environmental conditions. Conventional styles of stress evaluation, similar as tone- reported questionnaires, are frequently grounded on particular opinion and may not reflect real- time conditions more.

With advancements in machine literacy, it's now possible to dissect complex datasets and find patterns that may not be fact through conventional styles. This study focuses on prognosticating stress situations using machine literacy models trained on sleep- related and behavioral data, without depending on advanced physiological detectors.

II. LITERATURE REVIEW

Stress and sleep privation have been set up to be explosively identified in previous exploration. Reduced sleep length and poor sleep quality have been linked to increased situations of prostration, focus problems, and emotional insecurity, according to exploration.

Machine literacy has been extensively used in healthcare operations, similar as frazzle discovery and internal health vaticination. complicated relations in complicated datasets can be effectively covered by models like Random Forest and Gradient Boosting.

still, the maturity of current exploration uses big datasets or integrates physiological labels like brain exertion or heart rate variability. The foundation of this exploration is the lack of studies that concentrate on small datasets and are confined to behavioral and sleep- related characteristics.

III. METHODOLOGY

A. Dataset Description

The dataset applied in this study contains 60 participants and 14 variables covering different aspects of sleep, cognition, and lifestyle.

The variables include:

- Sleep duration (hours)
- Sleep quality score
- Daytime sleepiness
- Reaction time (Stroop Task, PVT)
- Memory performance (NBack accuracy)
- Emotional regulation score
- Age, gender, and BMI
- Caffeine consumption
- Physical activity level
- Stress level (target variable)

B. Artificial data creation

Because of the limited dataset size, artificial data creation was applied to expand the dataset. Two approaches were applied:

- Normal Distribution Sampling: Produces new values derived from the mean and standard deviation of every variable.
- Multivariate Sampling: Maintains connections among variables through covariance analysis, guaranteeing authentic data patterns.

These methods allowed for higher data availability while maintaining the original dataset's statistical integrity.

C. Machine Learning Models

The following models were applied:

- Linear Regression: Presumes a linear connection between input variables and stress levels.
- Random Forest: Utilizes various decision trees to enhance prediction precision.
- Gradient Boosting: Constructs models in a sequence to reduce prediction inaccuracies.
- XGBoost: A refined boosting algorithm recognized for its great efficiency and effectiveness

D. Evaluation Metrics

Model performance was tested using:

- R^2 (Deterministic Coefficient)
- RMSE (Root Mean Square Error)
- MAE (Mean Absolute Error)
- SMAPE (Symmetric Mean Absolute Percentage Error)
- Explained Variance

These metrics provide a comprehensive evaluation of prediction accuracy and error distribution.

IV. RESULTS AND ANALYSIS

A. Synthetic Data Validation

The synthetic dataset closely matched the original dataset in terms of mean and variability. This confirms that the generated data retained essential statistical properties required for model training.

B. Model Performance

The results showed varying performance across models:

- Linear Regression exhibited poor performance, suggesting that the connection between sleep and stress is not simply linear.
- Random Forest and Gradient Boosting demonstrated minor advancements but remained constrained in precision.
- XGBoost delivered the highest performance, exhibiting moderate prediction ability.

- Nonetheless, the overall performance of the model stayed quite low, indicating that variables related to sleep alone are insufficient for improved stress prediction
- Despite this, overall model performance remained relatively low, suggesting that sleep-related variables alone are not enough for better stress prediction.

C. Feature Importance

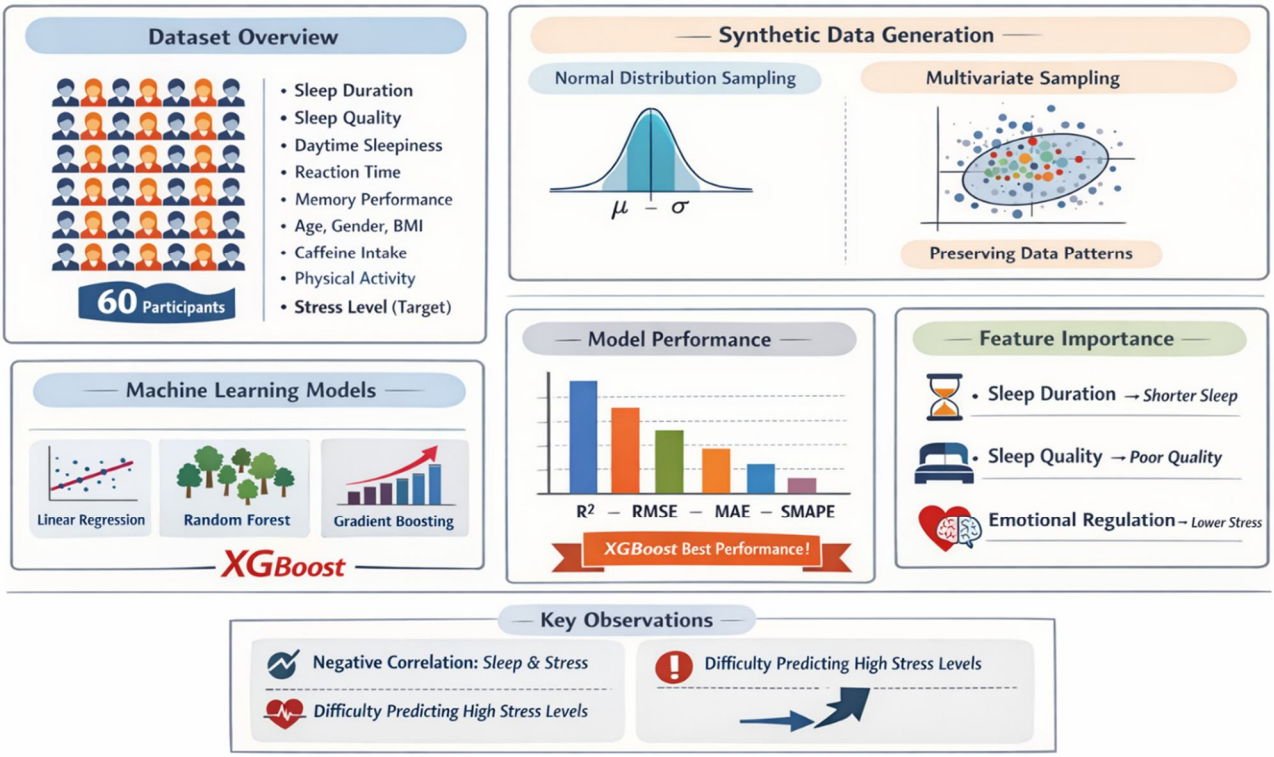
Key variables impact stress levels include:

- Sleep duration (shorter sleep associated with higher stress)
- Sleep quality (poor quality linked to higher stress)
- Emotional regulation (better regulation associated with lower stress)

D. Observations

Correlation analysis revealed a negative relationship between sleep duration and stress levels. However, the relationship was not strong enough to fully explain stress variability.

Prediction results also indicated that models struggled to capture extreme stress values and tended to predict values near the average.



V. DISCUSSION

The findings punctuate that stress is impact by multiple factors beyond sleep. While sleep is an important contributor, it alone can not completely explain variations in stress situations.

Synthetic data proved useful for perfecting model training but can not replace real- world diversity. The limitations observed in model performance emphasize the need for further comprehensive datasets.

VI. CONCLUSION

This study explored the use of machine literacy models to prognosticate stress situations grounded on sleep- related and behavioral data. While some connections were linked, the models showed limited prophetic delicacy.

Among the models tested, XGBoost performed stylish, but the results indicate that fresh data sources are necessary for meaningful stress vaticination.

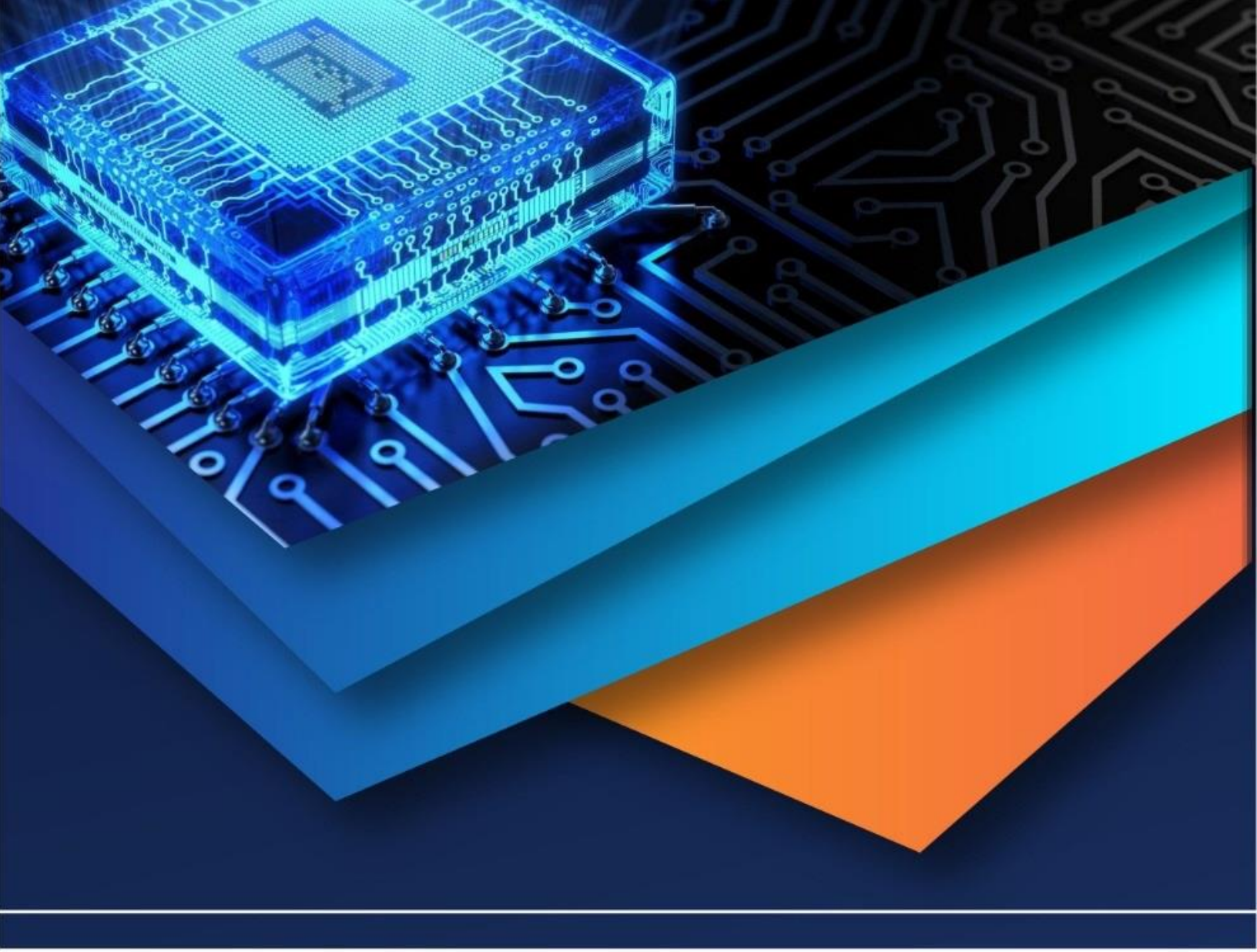
VII. FUTURE WORK

Future research directions include:

- Adding physiological data such as heart rate and EEG signals
- Collecting longitudinal data for better trend analysis
- Including environmental and psychological variables
- Using deep learning models for improved accuracy

REFERENCES

[1] M. A. Grandner, "Sleep, Health, and Society," Sleep Medicine Clinics, vol. 12, no. 1, pp. 1–22, 2017.
[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[3] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[4] J. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[5] C. A. Espie, "Insomnia: Conceptual Issues in the Development, Persistence, and Treatment of Sleep Disorder in Adults," Annual Review of Psychology, vol. 53, pp. 215–243, 2002.
[6] American Psychological Association, "Stress in America Survey," 2021.
[7] S. Cohen, D. Janicki-Deverts, and G. E. Miller, "Psychological Stress and Disease," JAMA, vol. 298, no. 14, pp. 1685–1687, 2007.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY