Predicting Sleep Quality From Daily Activity Logs Using Machine Learning

Dean Sacoransky (20112296), Dustin Pulver (20106164)
Department of Electrical and Computer Engineering
Queen's University

December 10, 2022
ELEC872

***Abstract*—Sleep-related disorders are a critical health problem facing society. This work attempts to find a representation between daily activity patterns and sleep quality. We explore machine learning methods for predicting an individual's sleep quality based on their activity logs from the previous day. We propose the use of supervised learning classifiers in combination with data preprocessing techniques and hyperparameter tuning. The testing results using Nu Support Vector Classification on the activity log data demonstrated a classification F1 score of 92.0%. Through a feature importance calculation, we found that laying down and caffeine consumption were the most significant activities impacting sleep quality.**

I. Introduction

A. Background

Sleep quality is one of the greatest factors impacting an individual's short-term and long-term health [1]. Like nutrition and exercise, sleep has a significant impact on mental and physical health [1]. Sleep deprivation can affect an individual's ability to concentrate, perform tasks, and can increase the risk of various health problems [1]. Sleep is necessary for preserving a high quality of life, and thus, it is important to discover how our actions relate to our sleep. The emergence of wearable sensors for tracking physical activity and sleep facilitates the development of new artificial intelligence applications.

B. Problem Statement

Sleep-related medical attention requires the physical presence of a patient through checkups and assessments. People are often reluctant to seek medical attention due to laziness, denial, mobility constraints, and cost. Additionally, medical attention is often only pursued in the case of sleep disorders, rather than people seeking to understand why they might have good quality sleeps. Thus, sleep-related medical advice is often used as treatment rather than preventative advice. There is a need for passive, daily monitoring of sleep quality to inform individuals on why their actions result in good or poor sleep quality.

C. Solution Overview

Previous work has achieved daily human activity recognition through ubiquitous sensing and machine learning techniques [2], [3]. We can leverage the automated human activity logs, and the relationship between daily activity and sleep, to predict an individual's quality of sleep on a given day. This work proposes a classification system of sleep quality based on an individual's daily activities. Through ubiquitous sensing and machine learning algorithms, it is possible to passively predict an individual's quality of sleep. This will allow the person to gain insight into the types of actions that promote healthy and poor sleeping habits, without the need for a medical checkup.

II. Related Work

A. Sleep Quality Prediction Based on Daily Activities

Previous studies have used wearable sensor data with machine-learning models to predict sleep quality. Hidayat *et al.* [4] used wrist-wearable sensor data to monitor an individual's physical activity level. This data was passed through a k-nearest neighbor (KNN) classifier to predict the subject's sleep duration and sleep efficiency. Sathyanarayana *et al.* [5] used actigraphy data from wearable sensors to perform binary classification on sleep quality. Several models were compared, including logistic regression, multilayer perceptron, convolutional neural networks (CNN), recurrent neural networks, long short-term memory (LSTM), and time-batched LSTM. In this study, CNNs produced the best results of up to 94%. Sadeghi *et al.* [6] used electrocardiogram signals from an Empatica wristband to classify sleep as light, medium, or deep. They used a CNN model with basic techniques such as max pooling and dropout layers to achieve a result of up to 75%. Thi Phuoc Van *et al.* [7] presented a novel adaptive algorithm based on ensemble learning to predict sleep efficiency. Specifically, they built a global model based on ensemble learning with common features from all clients and then combined the global model with more personalized features for each individual.

B. Pittsburgh Sleep Quality Index

The Pittsburgh Sleep Quality Index (PSQI) was developed in 1989 for psychiatric research in order to classify sleep quality into two categories: "good" and "poor" [12]. The evaluation of "good" and "bad" sleep quality is based on the subjects' self-rating of seven sleeping factors (i.e., subjective sleep quality, sleep latency, habitual sleep efficiency, sleep duration, awakening, sleep medication consumption, and daytime functioning). The index uses scores ranging from 0 to 21. A score lower than "6" corresponds to good sleep quality, and higher than ”6" represents poor quality. Previous works [13]–[15] showed effective binary classification models based on the PSQI using logistic regressions.

III. Experiment Setup

A. Data Source

This work made use of the Multilevel Monitoring of Activity and Sleep in Healthy People (MMASH) dataset to perform a binary classification of sleep quality as ”good” or ”poor” according to the PSQI. The MMASH dataset provides 24 hours of continuous heartbeat data, accelerometry data, sleep quality information, physical activity, psychological characteristics, and activity logs for 21 subjects. The MMASH dataset was created to promote research that tests the relationship between physical activity, sleep quality, and psychological characteristics. Several other works have applied machine learning techniques to the MMASH dataset. Specifically, [9]–[11] made use of heart rate time-series data to predict the expected heart rate of an individual. Bitkina *et al.* [8] used actigraph data from the MMASH dataset to assess quality of sleep. They applied the Cole-Kripke signal processing algorithm to the actigraph data to differentiate between sleeping and non-sleeping states, and calculated features such as time spent in bed, sleep duration, number of awakenings, and duration of awakenings. Furthermore, they applied a Support Vector Machine, Naïve Bayes classifier, logistic regression, and KNN model to classify sleep as good or poor quality according to the PSQI. However, rather than using antigraph or heart rate data, we used activity log data which indicates the duration and time of day that an individual performed a certain activity. The different activity classes and their corresponding activity IDs are presented in the numbered list below.

1) Sleeping
2) Laying Down
3) Sitting
4) Light Movement
5) Medium Movement
6) Heavy Movement
7) Eating
8) Small Screen Usage
9) Large Screen Usage
10) Caffeinated Drink Consumption
11) Smoking
12) Alcohol Consumption

IV. Method

This section presents the approaches and pipelines used for data pre-processing, classification tasks, and model tuning, starting with data pre-processing in IV-A, classification in IV-B, and model tuning in IV-C. This work was implemented with $Python$ 3.9.12 in a $Google\ Colaboratory$ notebook using the $Sklearn$ package.

A. Data Pre-processing

We explored several approaches for pre-processing such as label encoding, data cleaning, feature generation, and feature selection. We extracted two subsets from the original MMASH dataset (i.e., Activity and Survey) to be used as input features and training labels for our final dataset. Refer to Figure 1 as you read through this section to visualize the blocks of the pre-processing pipeline.

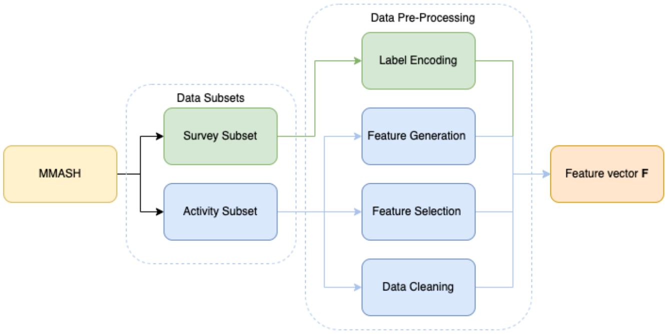

Fig. 1. Data-preprocessing system pipeline that implements label encoding on the output target (Survey Subset) and also implements data cleaning, feature generation, and feature selection on the input data (Activity Subset).

1) Label Encoding: The Survey subset includes the target feature of the proposed classification system, i.e., sleep quality according to the PSQI. During the 24-hour experimental period, each participant completed the PSQI questionnaire the morning after their sleeping period. The final PSQI value for each participant is an integer between 0 and 21, with values lower than 6 indicating good sleep quality, and higher than 6 indicating poor sleep quality [12]. Therefore, a suitable design choice was to map the PSQI ratings ranging from 0 to 21 to binary values that correspond to their respective sleep quality. We arbitrarily chose 1 to indicate good sleep quality and 0 to represent poor sleep quality. This concludes the first data prepossessing step, label encoding, which is necessary in order to perform the binary classification task.

2) Data Cleaning: The Activity subset of MMASH includes 12 activity categories reported from each user throughout the data collection period. Entries to the activity subset include the user identification, the start time and end time of the activity performed, day, and the activity identification. Some entries in the ’end time’ and ’activity class’ data fields contained null values. The corresponding data entries were removed from the dataset entirely, rather than performing data imputation. Data imputation is not a suitable choice in this scenario as inferring an ’end time’ or ’activity’ would not account for the relationships between the features.

3) Feature Generation: To better represent the daily activity of the participants, twelve new features, $\mathbf{Activity_n}$ (where n is the activity ID), were created to describe the total time spent by each user on each activity in seconds. A new data set was created which includes 21 participants’ user id $\mathbf{User_n}$ as row indices, 12 columns representing the total time spent performing each activity, and an output column displaying the corresponding sleep quality classification for each user. This data set will be referred to as $\mathbf{F}$ moving forward and can be seen in Table I.

TABLE I

THIS TABLE IS MEANT TO ILLUSTRATE, BY EXAMPLE, THE FEATURE VECTOR **F** AFTER DATA PRE-PROCESSING. THE INPUT DATA IS THE TOTAL TIME IN SECONDS THAT EACH USER PERFORMED EACH ACTIVITY, AND THE OUTPUT TARGET IS A BINARY LABEL OF SLEEP QUALITY ACCORDING TO THE PSQI.

User	$\mathbf{Activity_1}$[s]	$\mathbf{Activity_2}$[s] ...	$\mathbf{Activity_{11}}$[s]	**Sleep Quality**
$\mathbf{User_1}$	15420.0	780.0	300.0	0
$\mathbf{User_2}$	660.0	3120.0	800.0	1
$\mathbf{User_3}$	5400.0	108.0	0	0
...				
$\mathbf{User_{21}}$	7200.0	16800.0	240.0	0

4) Correlation-Based Feature Selection: To improve the performance of our final model, feature selection was conducted on the total activity data set to find the best subset of features. Variance-based feature elimination was conducted to discover the variance of each feature in the analyzed data set. A univariate variance close to 0 indicates that the feature does not contain significant information and thus should be dropped. All features in the data set were retained as they had large variance values. Correlation-based feature elimination was performed to determine the similarity between features and search for redundancy in the dataset. If two features are highly correlated, one can be removed as they essentially provide the same information. A correlation heat map of all input features in **F** can be seen in Figure 2. The heat map is labeled with the activity IDs for each feature, which can be found in Section III. Alcohol consumption and smoking activities, i.e., $\mathbf{Activity_{12}}$ and $\mathbf{Activity_{11}}$, had a correlation of 0.6 indicating strong correlation. These two behaviors can be logically linked to participants with negative habits. Smoking was eliminated from the feature set over alcohol as 17 of the participants did not smoke in the 24-hour period compared to only 8 participants that did not consume alcohol in the same time period. The remaining features were retained because no other pair of features had a high correlation value. Overall, the data pre-processing pipeline generated a feature set **F**, as seen in Table I, containing newly generated features describing the total time spent by each user to perform 11 activities. The target feature, Sleep Quality, was appended as a new column and encoded as binary values.

5) Dimensionality Reduction: Principle Component Analysis (PCA) was used as a method to generate a lower-dimension feature set. PCA is used mainly for dimensionality reduction, although the resultant principal components from PCA are ranked based on variance to produce a new feature set with the highest variance principal components. A parameter of PCA, $n\_components$, is used to specify the number of principal components in the resultant feature set, determined by the percentage of explained variance.

B. Classification Techniques

The dataset **F** used in this work consists of only 21 participant entries. Thus, we have elected to avoid deep learning methods and use techniques that have previously performed well on small datasets. The testing and training pipelines associated with the implementation of logistic regression, KNN, adaptive boosting, extreme gradient boosting, and support vector classifiers are explained in this section.

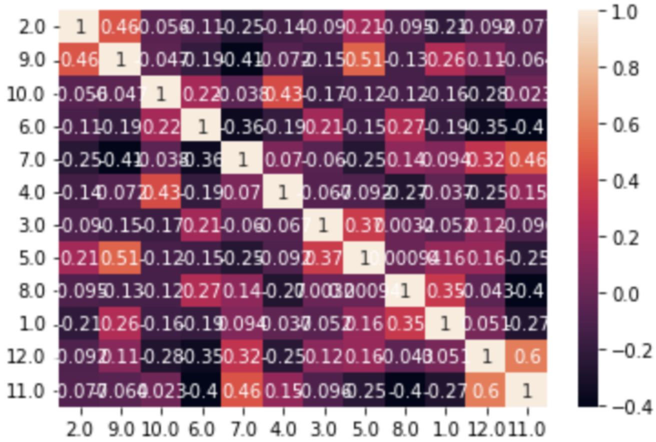

Fig. 2. Feature correlation heat map created with $Seaborn$ library, indicating a strong correlation between $\mathbf{Activity_{12}}$ and $\mathbf{Activity_{11}}$, i.e., alcohol consumption and smoking activities. This justifies our choice to eliminate the smoking feature. The activity names corresponding to each ID can be found in section III.

1) Logistic Regression: One approach to dealing with small data sets is using simple models to classify your data. Complex models tend to overfit with a lack of training data, thus logistic regression (LR) was selected as a suitable model for this classification task. For the tuning process, the hyperparameters include the regularization $penalty$ and c, the inverse of regularization strength, as seen in Table II.

2) KNN: Large datasets present the risk of noise which is known to adversely affect the performance of KNN. Our small dataset has no noise present and thus is a good fit for the KNN model. For the tuning process, the hyper parameters include $n\_neighbors$ and p, as seen in Table II. $n\_neighbors$ refers to the number of data points considered when determining the class label of a new data point. The hyper parameter p represents the function used to calculate the distance between a point and its neighbor. If p equals one Manhattan distance is used, if p equals two euclidean distance is used.

3) AdaBoost: AdaBoost or adaptive boosting is an ensemble boosting technique. Ensemble methods use the central limit theorem to produce low-variance models by averaging many high-variance models. Boosting is the process of creating multiple weak learners that add together to create a strong learner. With a small data set, every misclassified data point has a large impact on the accuracy of the model. In each iteration of the training process, AdaBoost gives more weight to data points that are misclassified in the previous iteration, creating a classifier that focuses more on misclassified data. This characteristic of AdaBoost makes it ideal to use on small datasets. For the tuning process, the hyper parameters include the number of estimators ($n\_estimators$) and $learning\_rate$.

TABLE II
SUMMARY OF THE MODEL-SPECIFIC HYPERPARAMETERS. A BAYSEIAN SEARCH ALGORITHM IS EMPLOYED OVER THE INDICATED RANGES TO FIND THE PARAMETER VALUE.

	Model Parameters			
Model	Parameter 1	range	Parameter 2	range
LR	c	[0.2-1]	$penalty$	[*l*1, *l*2]
KNN	$n\_neighbours$	[2-9]	p	[1,2]
Adaboost	$n\_estimators$	[20-50]	$learning\_rate$	[1e-4-1e-3]
XGBoost	$n\_estimators$	[10-100]	max_depth	[10-20]
SVC-Nu	nu	[0.2-0.5]	$Kernel$	[linear,poly,rbf]

4) XGBoost: XGBoost or extreme gradient boosting is another ensemble boosting method which utilizes a gradient boosting framework. Different than AdaBoost, gradient boosting treats the process of additively training and generating weak learners as a gradient descent algorithm over an objective function. The performance of XGBoost on small datasets has been shown to achieve improved performance compared to other models such as SVMs and ANNs [16]. For the tuning process, the hyper parameters include the number of estimators ($n\_estimators$) and maximum depth of the tree (max_depth).

5) Nu-SVC: Support vector classifier (SVC) maps data points to a high-dimensional space, finding the optimal hyperplane to split data points into two classes. Nu-SVC is mathematically equivalent to SVC although it introduces a new parameter nu. nu controls the number of support vectors and margin error. General SVMs perform well on small datasets, as their kernels require memory that exponentially scales with the given number of data points. For the tuning process, the hyper parameters include the kernel type used in the algorithm, $kernel$, and nu.

C. Model Tuning

We analyzed the impact of model tuning factors such as k-fold cross validation, PCA, and model-specific hyperparameters. The optimal set of model-specific hyper parameters was found through a Bayesian search, which samples over a distribution range for each parameter, as shown in Table II. General training hyper parameters were also explored, including the value of k for k-fold cross validation and $n\_components$ in the PCA algorithm. For each model, we perform an ablation study on PCA usage and vary the number of CV folds to understand the contribution of these components to the overall system, as seen in Tables III, IV, V, VI, and VII.

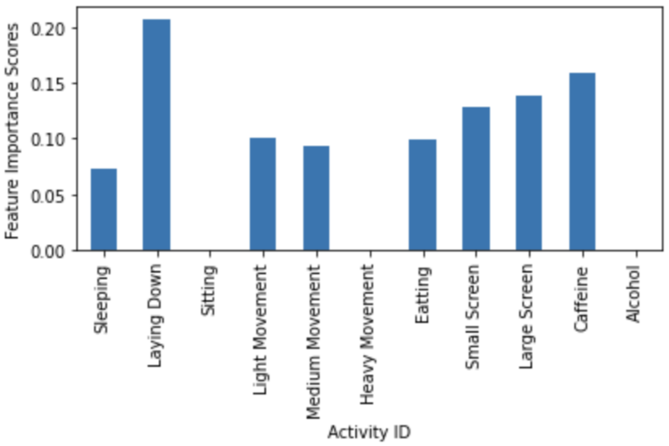

Fig. 3. Feature importance of activities for predicting sleep quality

TABLE III
PERFORMANCE OF LOGISTIC REGRESSION (LR) MODEL

LR	PCA		No PCA	
CV k-value	CV score	Test score	CV score	Test score
3	0.51	0	0.27	0.44
5	0.43	0.66	0.46	0.66
8	0.71	**0.80**	0.33	0.44

V. RESULTS

This section describes the results of the four described classification techniques, namely logistic regression, KNN, AdaBoost, XGBoost and Nu-SVC. The proposed classification system predicts whether each user will have a good or poor quality sleep based on their activities from the previous day. For each model, the test and cross validation $F1$ scores are presented. The best score for each model is bolded. We use the $F1$ score evaluation metric because the MMASH dataset is not perfectly balanced. As seen in Table III, logistic regression achieved a test score of 80% when utilizing PCA and 8-fold cross validation. PCA helped achieve a higher testing accuracy by finding features with higher variance. A higher k-value (number of folds) means that the model is trained on a larger training set and tested on a smaller test fold. This leads to a lower prediction error as the model learns from more of the data. The KNN model also yielded a test score of 80% when using 5-fold and 8-fold cross validation, as shown in Table IV. PCA had no impact on the result in this scenario. The AdaBoost model produced the worst results, a 54% test $F1$ score (Table V). XGBoost achieved a testing accuracy of 71% when using PCA (Table VI). Boosting methods are prone to overfitting when they put too much emphasis on correcting previously misclassified data. Finally, Nu-SVC achieved the highest test score of 92%. Support Vector classifiers are often extremely effective at classifying small datasets by maximizing the separation margin of classes in an altered feature space. Overall, PCA did not consistently work because high variance features are not necessarily important features.

Tree classifiers such as XGBoost provide the ability to calculate feature importance scores. The feature importance score of each activity class is presented in Figure 3. As expected, laying down and caffeine consumption were found to be the most significant features.

VI. CONCLUSION

Experimental analyses were conducted to determine the potential for using machine learning techniques to predict

TABLE IV
PERFORMANCE OF KNN MODEL

KNN	PCA		No PCA	
CV *k*-value	CV score	Test score	CV score	Test score
3	0.79	0.22	0.79	0.71
5	0.79	0.80	0.78	0.80
8	0.83	**0.80**	0.83	**0.80**

TABLE V
PERFORMANCE OF ADABOOST MODEL

AdaBoost	PCA		No PCA	
CV *k*-value	CV score	Test score	CV score	Test score
3	0.55	0.22	0.52	0.52
5	0.6	0.22	0.47	**0.54**
8	0.63	0.22	0.46	0.22

sleep quality of an individual based on their activity logs from the previous day. This work presented a pre-processing pipeline, model tuning techniques, and classification selection. The Nu-SVC model achieved the best performance of 92% $F1$ score when combined with either 8-fold cross-validation or 5-fold cross-validation and no PCA. Laying down and caffeine consumption were found to be the most significant activities impacting sleep quality through a tree-based feature importance analysis. This work contributes to the development of advanced sleep monitoring techniques that could be used in wearable sensing technology such as smart watches.

TABLE VI
PERFORMANCE OF XGBOOST MODEL

XGBoost	PCA		No PCA	
CV *k*-value	CV score	Test score	CV score	Test score
3	0.79	**0.71**	0.79	0.4
5	0.59	0.71	0.59	0.4
8	0.46	0.71	0.83	0.4

TABLE VII
PERFORMANCE OF NU-SVC MODEL

SVC-Nu	PCA		No PCA	
CV *k*-value	CV score	Test score	CV score	Test score
3	0.82	0.4	0.57	0.91
5	0.83	0.22	0.72	**0.92**
8	0.79	0.22	0.70	**0.92**

VII. STATEMENT OF CONTRIBUTION

The authors attest that they have spent an equal amount of time on this project.

REFERENCES

[1] Chattu VK, Manzar MD, Kumary S, Burman D, Spence DW, Pandi-Perumal SR. The Global Problem of Insufficient Sleep and Its Serious Public Health Implications. Healthcare (Basel). 2018 Dec 20.

[2] S.Ramasamy Ramamurthy and N.Roy, "Recent trends in machine learning for human activity recognition—a survey," Wiley Interdisciplinary Reviews: DMKD, vol. 8, no. 4, p. e1254, 2018.

[3] J.Wang, Y.Chen, S.Hao, X.Peng, and L.Hu, "Deep learning for sensor-based activity recognition: A survey," Pattern Recognition Letters, vol. 119, pp. 3–11, 2019.

[4] W. Hidayat, T. D. Tambunan and R. Budiawan, ”Empowering Wearable Sensor Generated Data to Predict Changes in Individual's Sleep Quality," 2018 6th International Conference on Information and Communication Technology (ICoICT), 2018, pp. 447-452.

[5] A. Sathyanarayana, S. Joty, L. Fernandez-Luque, F. Ofli, J. Srivastava, A. Elmagarmid, et al., ”Sleep Quality Prediction From Wearable Data Using Deep Learning", JMIR mHealth and uHealth, vol. 4, no. 4, pp. e125, 2016.

[6] R. Sadeghi, T. Banerjee and J. Hughes, ”Predicting Sleep Quality in Osteoporosis Patients Using Electronic Health Records and Heart Rate Variability," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020, pp. 5571-5574.

[7] N. T. P. Van, D. M. Son and K. Zettsu, ”A Personalized Adaptive Algorithm for Sleep Quality Prediction using Physiological and Environmental Sensing Data," 2021 8th NAFOSTED Conference on Information and Computer Science (NICS), 2021, pp. 113-119.

[8] O.V. Bitkina, J. Park, J. Kim, ”Modeling Sleep Quality Depending on Objective Actigraphic Indicators Based on Machine Learning Methods", 2022 Int J Environ Res Public Health, Aug 11-19.

[9] Oyeleye, M.; Chen, T.; Titarenko, S.; Antoniou, G. A Predictive Analysis of Heart Rates Using Machine Learning Techniques. Int. J. Environ. Res. Public Health 2022, 19, 2417.

[10] Site, A.; Lohan, E.S.; Jolanki, O.; Valkama, O.; Hernandez, R.R.; Latikka, R.; Alekseeva, D.; Vasudevan, S.; Afolaranmi, S.; Ometov, A.; et al. Managing Perceived Loneliness and Social-Isolation Levels for Older Adults: A Survey with Focus on Wearables-Based Solutions. Sensors 2022, 22, 1108.

[11] Geng, D.; Qin, Z.; Wang, J.; Gao, Z.; Zhao, N. Personalized recognition of wake/sleep state based on the combined shapelets and K-means algorithm. Biomed. Signal. Process. Control 2022, 71, 103132.

[12] Buysse, D.J.; Reynolds, C.F., 3rd; Monk, T.H.; Berman, S.R.; Kupfer, D.J. The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. Psychiatry Res. 1989, 28, 193–213.

[13] Jahrami, H.; BaHammam, A.S.; AlGahtani, H.; Ebrahim, A.; Faris, M.; AlEid, K.; Saif, Z.; Haji, E.; Dhahi, A.; Marzooq, H.; et al. The examination of sleep quality for frontline healthcare workers during the outbreak of COVID-19. Sleep Breath. 2021, 25, 503–511.

[14] Dzierzewski, J.M.; Mitchell, M.; Rodriguez, J.C.; Fung, C.H.; Jouldjian, S.; Alessi, C.A.; Martin, J.L. Patterns and predictors of sleep quality before, during, and after hospitalization in older adults. J. Clin. Sleep Med. 2015, 11, 45–51.

[15] Seun-Fadipe, C.T.; Mosaku, K.S. Sleep quality and academic performance among Nigerian undergraduate students. J. Syst. Integr. Neurosci. 2017, 3, 1–6.

[16] Zou M, Jiang WG, Qin QH, Liu YC, Li ML. Optimized XGBoost Model with Small Dataset for Predicting Relative Density of Ti-6Al-4V Parts Manufactured by Selective Laser Melting. Materials (Basel). 2022 Aug 1.