Establishing the Correlation Between Bio-Data and Concentration Level Utilizing EEG Data

Seojeong Park[1], Seungah Jang[2], Jeongmin Seo[3], Yunhui Lim[4], Byeongsoo Min[4], and Martin Kim[5]

[1]Computer Science, Hallym University
Chuncheon, South Korea
diditjwjd@gmail.com
[2]Computer Science and Statistics, Jeju National University
Jeju, South Korea
jang6733@jejunu.ac.kr
[3]Software Convergence, Kyunghee University
Seoul, South Korea
balljm@naver.com
[4]Computer Science and Engineering, Kyunghee University
Seoul, South Korea
lyhhh0320@gmail.com, qud9783@gmail.com
[5]Computer and Information Technology, Purdue University
West Lafayette, USA
19alswprla@gmail.com

***Abstract*—In modern society, the management of self-concentration has gained significance. It is essential to propose a method that enables concentration monitoring in daily life. This study aims to establish a correlation between bio-data from a wearable device and concentration by performing regression models. Bio-data is directly collected and subjected to a six-stage of preprocessing process. The preprocessed dataset selected for training models comprises six types of bio-data: Heart Rate (HR), Heart Rate Variability (HRV), Body Movement, Sleep, Electrodermal Activity (EDA), and Wrist Temperature. The model exhibiting the highest performance is the optimized Extra Tree Regressor, which achieves an R-squared of 0.876. The study indicates that lower EDA significantly enhances concentration, showing a feature importance of 0.23, and identifies a proportional relationship between Body Movement, Sleep, and concentration levels, each with a feature importance of 0.19. Therefore, the objective indicator is applied as the concentration score by analyzing bio-data. The concentration score calculated based on the model is provided to user through the application.**

***Index Terms*—concentration, bio-data, EEG, wearable device, Extra Tree Regressor**

I. Introduction

Modern day society has bore witness to a phenomenon of diminished concentration levels and attention spans as a result of dependency and increased usage of smart phones. The usage rate of smartphones is more than 70% of adults in the United States and nearly 50% of adults worldwide. The decline in cognitive ability in daily life due to the increase in smartphone use can be observed in various forms. Media multitasking has reduced memory work ability as well as lower sleep quality, leading to a decrease in cognitive ability. Moreover, anxiety symptoms caused by inability to use smartphones negatively affect cognitive function [1].

Therefore, the need for methods to facilitate and assist users in maintaining high levels of concentration in their daily lives has increased significantly. Research on electroencephalography analysis is being actively conducted in the field of brain science related to concentration. When a person performs a specific activity or feels an emotion, electrodes are generated in the brain. These electrodes are observed in the form of waves and are called electroencephalography (EEG). Based on their frequencies, the EEG is divided into five waves of α, β, δ, θ and γ waves. β waves are released when the brain is concentrating on a task or is cognitively awake. θ waves erupt in a drowsy state or a physically relaxing situation [2].

There is a strong correlation between biological data (bio-data) and EEG. Changes in bio-data can be observed to correspond with alterations in specific brainwaves within the EEG. Further details on this relationship are discussed in Section 2. Moreover, while bio-data is associated with concentration, there has been limited research aimed at deriving a clear correlation between comprehensive bio-data information and concentration levels. The objective of this study is to analyze various bio-data to indicate concentration levels, thereby deriving the correlation between them with importance and proportionality.

In this study, a total of nine machine learning models are selected: Extra Tree Regressor, RandomForest Regressor, XGBoost Regressor, etc. These models are employed to analyze bio-data and brainwaves collected through the wearable device and the EEG sensor in order to quantify concentration level. They undergo performance evaluation, and comparative analysis to determine their effectiveness in predicting concentration level.

The remainder of this paper is organized as follows: Section 2 provides a literature review; Section 3 illustrates methodology for the experiment; Section 4 demonstrates the result of model performance and feature importance; Section 5 describes the implementation for the service, and Section 6 explains the conclusion.

II. Literature Review

Numerous prior studies have explored the correlation between EEG and concentration, the relationship between bio-data and EEG, as well as the connection between bio-data and concentration.

A. Brain Signal Analysis for Concentration Level Prediction

Several papers have used the analysis of brain signals to predict concentration levels. Tae Jin Choi *et al*. [3] utilized the formula of [β + SMR / θ]for concentration evaluation by applying a Fourier transform (FFT) to raw EEG data. The result indicated an increase in β waves and a decrease in θ waves during focused states. Ning-Han Liu *et al*. [4] carried out a concentration experiment where participants answered questions using the MindSet brainwave sensor. Through FFT on raw EEG data, they extracted α, β, δ, θ, and γ waves. The concentration evaluation metric, R index = (α/β), combined human-analyzed concentration states and achieved 76.82% accuracy using SVM model. Varsha T. Lokare *et al*. [5] utilized the Muse headband to assess concentration during everyday repetitive tasks. They used α, β, δ, θ and γ waves as inputs, incorporating self-assessment to derive four categories of concentration level. Machine learning models including Naive Bayes, ANN, SVM, and Decision Tree were applied, with ANN showing the highest accuracy of 71.46%.

Upon reviewing studies analyzing concentration-related brain signals, two distinct methods for deriving concentration levels were evident. One involved using a mathematical formula, while the other utilized self-assessment. The former necessitates specialized brainwave sensors capable of providing raw EEG data. The latter, while using more affordable EEG sensors, provided subjectivity in concentration assessment metrics. This study aims to take the advantages of both methods, allowing for the utilization of affordable EEG sensors to establish an objective concentration assessment metric.

B. Objective Evaluation Metric

C. Hasegawa *et al*. [6] introduced the R index as a concentration assessment metric. R = (α/β), where diminishing R values indicate a larger proportion of β waves, signifying a more alert state of the brain. Lutsyuk, N.V. *et al*. [7] proposed four concentration assessment metrics: θ Spectral Powers(SP), α SP, β SP, and β/θ SP. Pearson correlation coefficient analysis displayed the highest positive correlation with β/θ SP during concentrated brain states.

C. Correlation Between Bio-data and Brain Signals

Prior studies examined the relationship between bio-data and brain signals. Ji-Won Kwon *et al*. [8] simultaneously analyzed EEG and heart rate variability to assess concentration, stress, and tension among surgeons using wearable devices during surgery. Pearson correlation analysis revealed a positive correlation between concentration and β waves, and a negative correlation with θ waves. Martínez Vásquez *et al*. [9] analyzed electrodermal(EDA) and EEG from participants performing four cognitive tasks, indicating a high mutual information value between EDA and the δ component. Kim SC *et al*. [10], observing EEG in neurological regions, revealed that insufficient sleep resulting in low α waves leads to inattentive concentration. These studies have shown the correlation between bio-data and certain brain waves through experiments conducted in specific scenarios.

D. Relationship Between Bio-data and Concentration

Numerous studies have explored the relation between bio-data and concentration. Alba G *et al*. [11] indicated that higher HRV was associated with better performance in various cognitive tasks related to attention, working memory, and inhibitory control. Taraneh Aminosharieh Najafi *et al*. [12] demonstrated that EDA serves as an indicator of both concentration and arousal states. Hans P.A. Van Dongen *et al*. [13] highlighted significant cumulative performance deficits in tasks measuring alertness, working memory, and cognitive throughput among individuals with insufficient sleep duration.

Based on the facts presented in prior research, this study aims to provide a clear correlation for deriving concentration level from bio-data. Objective concentration evaluation metrics were employed instead of self-assessment. Pre-extracted five brainwaves were used without Fourier transform preprocessing. Focusing on everyday concentration rather than assuming specific task scenarios, this study collected bio-data from wearable devices that can easily be worn in daily life, avoiding the use of specialized brainwave sensors.

III. Methodology

In this section, the research focuses on analyzing the impact of bio-data obtained from wearable devices on brain waves, specifically its correlation with concentration levels and the methods to quantify this influence. This part includes preprocessing steps for datasets used in algorithms and the exploration of effective machine-learning models to predict concentration levels.

A. Data Collection and Experiment Conditions

The experiment was conducted on 6 students (3 males, 3 females), between ages 22 - 25 over eight weeks. Each subject conducted 6 experiments per day, and each test took 20 minutes. To avoid experimental bias on each test, a minimum interval between each test was set over 30 minutes. The participants were asked to simultaneously wear two different devices on their body: a Fitbit, a wearable device on their wrist, and an EEG sensor device on their head.

The subjects were not restricted to any temporal or spatial limitations, and they were asked to engage in various tasks such as studying, watching media content, exercising, and reading papers. Additionally, the subjects performed a concentration game at least once a day in order to obtain experimental data during a high-concentration period. Tests were mainly conducted in calm and uncrowded spaces but some were carried out in noisy atmospheres, both of which are common in daily life.

B. Measuring Bio-Data

Two devices are used in this study. First, Fitbit Sense 2 collects Active Zone Minutes (AZM), EDA, Deep Sleep in Minutes, and Wrist Temperature. Second, the Flowtime Biosensing Meditation Headband collects Attention Score and raw data on Brainwave, HRV, HR, and Coherence flag value. Data is collected every 0.6 seconds during measurement.

1. Fitbit Sense 2 (Fitbit)

Active Zone Minutes (AZM): AZM are measured as two values when activity is detected. It consists of the FAT_BURN and CARDIO properties. AZM is measured using the heart rate zones. It is recorded every minute in the activity zone. FAT_BURN is when the heart rate is in a low heart rate zone and is defined as 1 minute. CARDIO is when the heart rate is in the vigorous/maximal heart rate zone and is defined as 2 minutes. In this study, the time that is not in the activity zone is defined as NORMAL, which has a value of 0.

EDA: EDA refers to the variation of the electrical properties of the skin in response to sweat secretion [14]. EDA data is Electrodermal activity level.

Deep Sleep in Minutes (Sleep): Sleep consists of four stages: Awake, REM, Light, and Deep Sleep. Deep sleep promotes physical recovery and aspects of memory and learning [15]. Deep Sleep in Minutes represents the sleep time in minutes in the Deep Sleep stage.

Wrist Temperature: Skin temperature is the temperature of the skin's surface [16]. It represents a change of degree in temperature. It has a positive value or a negative value.

2. Flowtime Biosensing Meditation Headband (Flowtime Headband)

Brainwave: Brainwaves are produced by synchronized electrical pulses from masses of neurons communicating with each other. [17] Brainwaves are classified into five types according to frequency : α wave, β wave , θ wave, δ wave, and γ wave. The range of value is [0, ∞).

Attention Score: The device collects real-time brainwaves and measures the user's attention level. It is calculated by analyzing the spectral characteristics of brainwaves [18]. The range of value is [0, 100].

HRV: Heart rate variability (HRV) is the physiological phenomenon of variation in the time interval between heartbeats. The higher the value is, the greater your heart rate changes. The lower, the smaller [19]. The range of value is [0, 255].

HR: Heart Rate (HR) is the number of times a person's heartbeats per minute [20]. The range of value is [0, 255].

Coherence: Coherence indicates whether the body is in a coherent state. It has two values: 1 is in a coherent state, and 0 is in an incoherent state.

C. Concentration Indicator Selection

In the case of the attention score provided by the Flowtime Headband, it is related to the special case of meditation. Therefore, this study, which targets daily life situations, had to select another reasonable concentration indicator based on the literature review.

Grounded in the literature review, this study has identified two potential concentration indicators derivable from the possessed brain wave data: the β/θ Spectral Powers and the R index. Among these, the R index exhibits an inverse relationship with concentration scores, prompting a comparative analysis between the Reverse R index and the β/θ Spectral Powers in this research. Both indicators demonstrated a linear relationship with the attention scores yielded from Flowtime Headband. Consequently, the Pearson Correlation Coefficient, a metric quantifying the linear correlation between two variables, is employed to select an indicator of concentration.

Pearson Correlation Coefficient is defined as the covariance of the two variables divided by the product of their standard deviations. Governed by the Cauchy-Schwarz inequality, this coefficient assumes values between +1 and -1. A coefficient of +1 signifies a perfect positive linear correlation, 0 indicates no linear correlation, and -1 represents a perfect negative linear correlation [21].

Upon comparing the Pearson Correlation across 3,435 EEG data, the correlation coefficient between β/θ Spectral Powers and the attention score is found to be 0.750154160664315. In contrast, the Pearson Correlation coefficient between the Reverse R Index and the attention score is 0.7205323406010573. Consequently, this study selects β/θ Spectral Powers as the concentration indicator.

D. Data Preprocessing

This study undergoes a six-stage data preprocessing process, which includes augmentation, error removal, feature engineering, imputation, feature scaling and feature selection.

1. Augmentation

TABLE I
R-SQUARED SCORE ACCORDING TO TIME INTERVAL SETTING

Time interval	Samples	R^2 score
(1) 10 seconds	5,763	0.8642836913
(2) 12 seconds	4,798	0.8557038406
(3) 15 seconds	3,825	0.8283565023
(4) 20 seconds	2,867	0.7817081704

In this study, data is extracted from two different devices, each having distinct time intervals. The Fitbit outputs data in one-minute intervals, while the Flowtime Headband outputs data approximately every 0.6 to 0.7 seconds. Therefore, a process to unify the time intervals and match timestamps between these datasets is essential.

To set an appropriate time interval, Fitbit data undergoes interpolation using the linear interpolation. For the Flowtime Headband data, it is grouped by time interval, with the mean value set as the representative value for each group. Linear Interpolation is a method of curve fitting using linear polynomials to construct new data points within the range of a discrete set of known data points [22].

Furthermore, it is observed that setting a smaller time interval results in the larger quantity of data. In this study, to prevent distortion in bio-data, the use of data augmentation techniques is avoided. Therefore, the study increases the amount of data by adjusting the time interval of its own dataset.

Table I presents the R-squared values and the size of each dataset when evaluated with the final model, corresponding to different time interval settings. When the dataset is set to a 10 seconds time interval, it exhibits the largest dataset size and the best R-squared value.

2. Error Removal

TABLE II
EUCLIDEAN DISTANCE - CASE WITH FEW ERRORS

	Remove errors and calculate mean in time interval group	Not removing errors and calculate 15% trimmed mean in time interval group
alpha wave	0.003682500000010691	0.3134747614525308
hr raw data	0.000000000000000000	2.2528755895858237

TABLE III
EUCLIDEAN DISTANCE - CASE WITH MANY ERRORS

	Remove errors and calculate mean in time interval group	Not removing errors and calculate 15% trimmed mean in time interval group
alpha wave	2.10012115868947	0.32638866880899
hr raw data	122.11822529274264	19.39874768394027

During data collection, instances of errors in the data from the Flowtime Headband are observed. These errors include cases where brain waves are continuously uniform and instances where the hr value is zero.

To address these errors, this study compares two methods. The primary method involves grouping data by time interval, followed by detecting and addressing errors within these groups. For this, a criterion time within the time interval group is necessary to determine if the group is an error. In this study, this criterion time is set to range from 70% to 80% of the time interval duration. If the duration of errors within a time interval group exceeds criterion time, the group is treated as missing value. Conversely, if the duration of errors is less, the group's representative value is set as the average of the remaining time after removing errors. For example, if the time interval is 10 seconds and the criterion time within the group is 7 seconds, a group with errors occurring for 8 seconds will be replaced with missing value. However, if errors occur for 4 seconds within a 10 seconds group, the representative value of that group will be the average of the remaining 6 seconds.

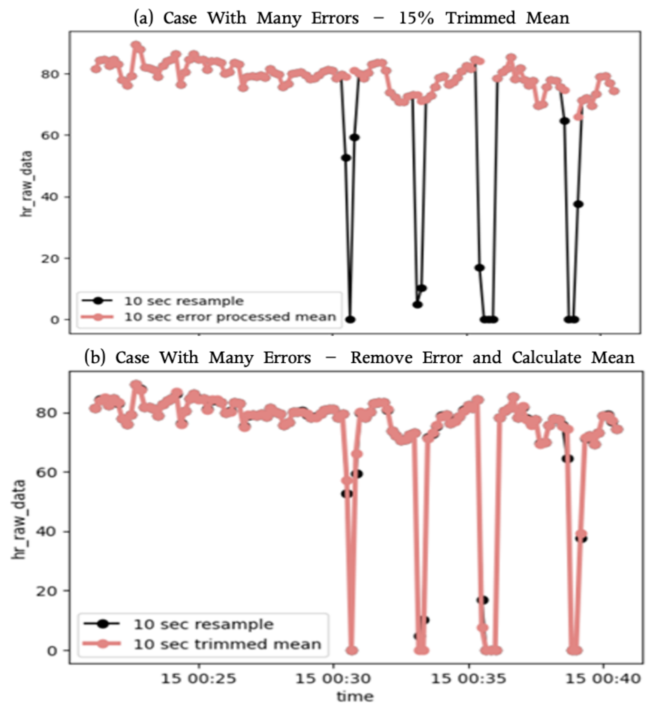

Fig. 1. Comparing Two Error Removal Methods - Case With Many Errors

The second method for addressing errors is grouping data by time interval and then employing trimmed mean for error elimination. In other words, trimmed mean is used to detect and remove outliers representing errors.

To choose between these two methods, the study measures the Euclidean distance between each method and the original Flowtime Headband data that is resampled at 10-second intervals. As indicated in Table II, during experiments with minimal errors, the first method shows a closer resemblance to the original data compared to the second method, effectively preserving the original data in cases with almost no errors. On the other hand, as shown in Table III, during periods with frequent errors, the first method demonstrates a longer distance from the original data compared to the second method, indicating successful error handling by the first method. Graphical analysis, as seen in Fig. 1 (b), reveals that the second method tends to follow the errors of the original data. However, as illustrated in Fig. 1 (a), the first method does not follow the original data with errors. Therefore, in this study, the first method is used, as it better preserves the trend of the original data in error-free cases and successfully handles errors in cases with errors.

3. Feature Engineering

The majority of the experiments in this study are conducted

TABLE IV
R-SQUARED SCORE BY CUMULATIVE ACTIVE ZONE MINUTES

Time period of cumulative sum	R^2 score
1 hour	0.83972
1 hour 30 minutes	0.85472
2 hours	0.86239
2 hours 30 minutes	0.86296
3 hours	**0.86428**

TABLE V
MSE VALUES COMPARED TO DATASET WITHOUT MISSING VALUES

	Wrist Temperature	EDA
Iterative Imputation	2.8693930001028156	24.336120649736955
Imputation using Random Forest Regressor	**0.017327493178123897**	**0.5766367293404618**
K-NN Imputation	0.19606558047255002	144.00392573470927

in non-exercising conditions. As a result, the Active Zone Minutes value in the collected data is predominantly zero, making it difficult to derive meaningful results from this feature. Based on the premise that cognitive functions, such as attention and concentration, are improved after exercise [23], a new feature is developed in this study. This feature, named 'Body Movement,' is quantified as the cumulative sum of Active Zone Minutes within a specified time period prior to the experimental session. As demonstrated in Table IV, the aggregation of 3 hours of Active Zone Minutes resulted in the highest R-squared score.

4. Imputation

After the previous preprocessing steps, the total number of rows in the dataset is 18,653. However, dropping rows with missing values reduces the dataset size to 5,619. To address the significant reduction in data volume due to missing values, imputation techniques are employed. Data imputation is a method that retains the majority of a dataset's data and information by replacing missing data with alternative values [24]. In this study, the imputation techniques utilized include Iterative Imputation, Imputation using Random Forest Regressor, and K-NN Imputation.

Iterative imputation fills in missing values by generating plausible numbers derived from distributions of and relationships among observed variables in the data set [25]. Random forest is a commonly-used machine learning algorithm, which combines the output of multiple decision trees to reach a single result [26]. K-NN (k-nearest neighbors) is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point [27].

Furthermore, imputation is conducted on Wrist Temperature and EDA, which have the highest incidences of missing values.

To select one of these three techniques, missing values are randomly created for a dataset without missing values, followed by imputation. As shown in Table V, the Imputation using a Random Forest Regressor demonstrates the smallest Mean Squared Error (MSE). Therefore, employing the Random Forest Regressor for data imputation, which shows the smallest MSE value, the data volume increases to 12,358 rows.

5. Feature Scaling

In order for machine learning models to interpret features on the same scale, this study conducts feature scaling [28]. The methods employed for feature scaling are standardization and min-max normalization. For the Wrist Temperature feature, which represents the rate of temperature change and includes negative values, standardization is utilized to reflect this significance.

Other features, not exhibiting normal distribution characteristics, are subjected to min-max normalization. Additionally, as this research utilizes tree-based models, it is appropriate to apply different scaling techniques to individual features.

6. Feature Selection

From the two devices, data is obtained in seven different categories. The final dataset is constructed using the best combination of features, which are selected through several experiments involving various feature combinations. Based on previous studies [29], EDA, Wrist Temperature, and HRV are included as essential components in these combinations. Extra Tree Regressor is used as the machine learning algorithm for comparing each dataset. As a result, the feature combination of HR, HRV, Body Movement, Sleep, EDA, and Wrist Temperature demonstrated the best performance with an R-squared score of 0.8707 in predicting the SP ratio, which is the primary focus of our research.

Ultimately, Table VI displays the result of the entire preprocessing process. A progressive increase in performance is observed according to each preprocessing step.

E. Machine Learning Models

1. Model Selection

The process of selecting a regression algorithm for use in the research involved investigating various machine learning models for a regression task using bio-data. This study examined the performance of different models. The dataset used in the model selection process is prepared in the previous stage, and the preprocessed dataset contains 12,358 rows. The entire dataset is divided into train and test sets with a ratio of 7:3, and experiments are conducted using k-fold cross-validation with 5 folds to ensure that the models could observe diverse training data. Additionally, all models are utilized with default parameters and are not optimized.

As shown in Table VII, the Extremely Randomized Tree Regressor exhibited the highest performance with an R-Square of 0.858. Following closely are the RandomForest Regressor and XGBoost Regressor. These models are tree-based ensemble models known for their ability to capture patterns in nonlinear data. The Extremely Randomized Tree Regressor and RandomForest Regressor utilize bagging, while the XGBoost Tree Regressor employs boosting for ensemble learning.

dataset	samples	Extra Tree Regressor	RandomForest Regressor	XGBoost Regressor
(1) Original	1012	0.6208360258	0.5751215225	0.5425420553
(2) Augmentation	5763	0.8245028924	0.7815207115	0.7473824518
(3) Removing Errors	5619	0.8081339869	0.7860800542	0.744537729
(4) Feature Engineering	5619	0.8399870882	0.7909246149	0.780730066
(5) Imputation	12358	0.8630560284	0.8374415757	0.796647765
(6) Feature Scaling	12358	0.8642836913	0.8342062226	0.8029168307
(7) Feature Selection	12358	0.8707710654	0.8393843648	0.8014808609

TABLE VI
MODEL RESULTS DEPENDING ON EACH PREPROCESSING STEP WITH R-SQUARED SCORE.

TABLE VII
MODEL RESULTS WITH SELECTED DATASET

Algorithms	R-Squared
Extra Tree Regressor	**0.8583776593**
RandomForest Regressor	**0.8267849591**
XGB Regressor	**0.7900022036**
LGBM Regressor	0.7578767966
KNN Regressor	0.6992657885
Decision Tree Regressor	0.6745135311
Gradient Boost Regressor	0.5505307193
Support Vector Regressor	0.3748954896
AdaBoost Regressor	0.3341737998

2. Model Optimization

This is a brief description and operational mechanism of the three models that exhibited the best performance.

a) RandomForest Regressor: The RandomForest Regressor [30] was initially introduced in a paper by the American statistician Leo Breiman in 2001. It employs a bagging ensemble learning method based on multiple Decision Trees. The algorithm randomly composes several sample datasets and combines the results of decision trees based on each dataset to achieve stable and powerful predictions.

b) Extremely Randomized Tree (Extra Tree Regressor): The Extremely Randomized Tree algorithm [31] focuses on randomness. Similar to the working principle of RandomForest, it constructs decision trees based not only on random combinations of data samples but also randomly selects features for node splitting during the decision tree construction process.

c) Extreme Gradient Boosting Regressor (XGB Regressor): The Extreme Gradient Boosting Regressor [32], as the name suggests, utilized gradient boosting instead of bagging. It is an ensemble algorithm where multiple decision trees learn from the errors of the previous models using gradients. Unlike the basic GBM algorithm, the XGB Regressor implements parallel processing, leading to faster learning speeds.

Following the identification of the top-performing three models, optimization is conducted. Utilizing the hyperparameter optimization tool provided by the Scikit-learn library, GridSearchCV, various combinations of hyperparameters are explored to enhance model performance and determine the optimal hyperparameter settings.

3. K-Fold Cross-Validation

In the evaluation of model performance, it is essential to distinguish between training and evaluation data. The direction in which the model learns can significantly differ based on the distribution and ratio of training and validation data, exerting a substantial influence on the model's overall performance. To address this, a validation dataset was introduced to assess the training data, employing cross-validation algorithms for diverse data combinations during model training.

Considering the dataset's size, different cross-validation folds are experimented with, and found that 5-fold and 10-fold provided optimal descriptions of the data. The results of each cross-validation are detailed in Table 8. With a 5-fold configuration, 6,623 sample data points are used for each training iteration, while a 10-fold setup utilized 7,451 samples for model training.

4. Evaluation Metrics

The three metrics used for performance evaluation are as follows: RMSE, MAE, R-Sqaured.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_j - \hat{y}_j)^2} \tag{1}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2 \tag{2}$$

$$R^2 Score = 1 - \frac{\sum_{i=1}^{n}(y - \hat{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2} \tag{3}$$

Root Mean Squared Error (RMSE) is calculated as the square root of the average of the squared error for all samples. Squaring the errors assigns greater weight to larger errors, making RMSE sensitive to outliers. Additionally, taking the square root of the mean of squared errors ensures that the unit of measurement is consistent with the original values, providing an intuitive interpretation. As RMSE represents the mean of errors, a smaller value indicates a better-performing model.

Similarly, Mean Absolute Error (MAE) represents the average error across all samples. Unlike RMSE, MAE uses absolute values, allowing for easy interpretation as the unit of measurement remains consistent with the output values. MAE is less influenced by outliers since it does not involve squaring the errors. It is often used alongside RMSE when comparing performance based on errors.

The R^2 Score (R-Squared) is a metric related to how well a regression model explains the variability in the dependent variable. It is calculated by considering the ratio of the sum of squared differences between the predicted values and the actual values to the sum of squared differences between the actual values and their mean. A value closer to 1 indicates that the model effectively explains the data, while values closer to 0 suggest poor explanatory power. In the case of a negative value, it implies that the model performs worse than predicting the mean level of the data.

5. Post-processing

The process of scaling the concentration scores derived from the model results into a range the user can understand is necessary. Initially, during data preprocessing and model training, the dependent variable, SP ratio, is scaled to values between 0 and 1. Consequently, the model's predicted results also need to be transformed into the [0, 1] range. To achieve this, a Clipping process is applied, replacing any results outside the [0, 1] range with the respective minimum values. Subsequently, the values are MinMaxScaled into the commonly used score range of [0, 100]. The final scaling allows for the derivation of scores that will be presented to the users.

IV. Result

A. Result Analysis

TABLE VIII
MODEL EXPERIMENT RESULTS

Model	k=5			k=10		
	R^2	RMSE	MAE	R^2	RMSE	MAE
ETR	0.8600	0.0902	0.0609	0.8745	0.0854	0.0579
Optimized ETR	0.8610	0.0899	0.0608	0.8763	0.0848	0.0576
RFR	0.8286	0.0998	0.0680	0.8430	0.0955	0.0652
Optimized RFR	0.8316	0.0989	0.0674	0.8444	0.0951	0.0649
XGBR	0.7979	0.1084	0.0780	0.7993	0.1084	0.0774
Optimized XGBR	0.8270	0.1003	0.0701	0.8314	0.0990	0.0689

The results obtained through the described process are presented in Table VIII. Upon careful examination of the table, the following observations are made. Setting 10 folds instead of 5 resulted in an average performance improvement of 0.01. Although the differences due to optimization are insignificant, there is noticeable progress. Ultimately, the Optimized Extra Tree Regressor model exhibited the highest performance with an R-squared of 0.876 in the 10-fold learning environment, leading to its selection as our final model.

B. Feature Importance

This study examined the impact of each feature on the predictions using feature importances in Fig. 2. For the Extra Tree Regressor model, which demonstrated the highest performance, the influential features in descending order are found to be EDA, Wrist Temperature, Body Movement, and Sleep.

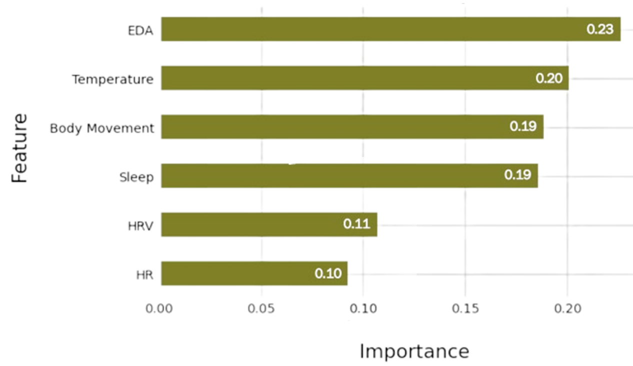

Fig. 2. Feature Importance Analysis.

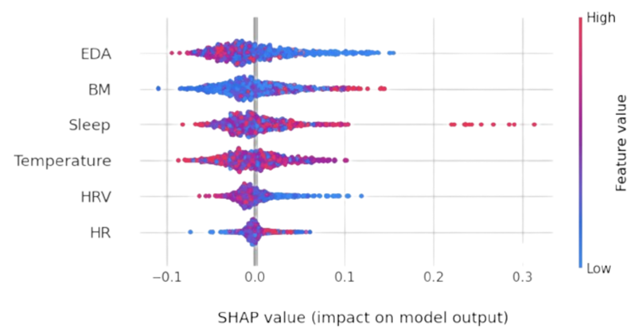

Fig. 3. Feature Importance Analysis using SHAP.

Explainable AI, known as XAI, has become widely utilized today as a tool to interpret black-box models that have nontransparent internal structures and are challenging to interpret easily, particularly in deep learning and tree-based machine learning models. Among these tools, SHAP [33] (SHapley Additive exPlanations), developed by Microsoft Research researchers Scott Lundberg and Su-In Lee in 2017, is a prominent XAI method. This is based on the concept of Shapley values originating from game theory and is employed to assess the contribution of each feature to the model's output. Using SHAP to examine the contributions of each feature, the following insights are gained.

Each sample is represented in blue for small feature values and in red for large feature values in Fig. 3. Observing how the prediction values change in relation to the magnitudes of features provides insights into the correlation between each feature and the predicted outcome, which is the concentration metric. EDA influences an increase in concentration when its value is small. Therefore, EDA can be considered inversely proportional to concentration; as EDA decreases, concentration tends to incease. It can be explained by the correlation between stress levels and EDA [34]. As stress levels rise, body undergoes automatic reactions, such as sweating secretion, leading to physiological changes. Consequently, an increase in skin conductance follows, manifesting an elevation in EDA. Considering this, it can be inferred that EDA has a positive correlation with stress. Lower stress values correlate with improved concentration. On the contrary, Body Movement and Sleep contribute to an increase in predictions. This suggests that adequate sleep and body movement prior to concentration tasks enhance overall concentration.

V. Implementation

In this paper, an application is developed that calculates the user's concentration score and provides the information which includes bio-data such as EDA, Sleep, HR, HRV, and Wrist Temperature for the user. As shown in Fig. 4, the Evaluating Concentration Level System Architecture that has been formulated can largely be divided into three sections based on its function. It consists of a front-end section, a back-end section, and a machine learning model section.

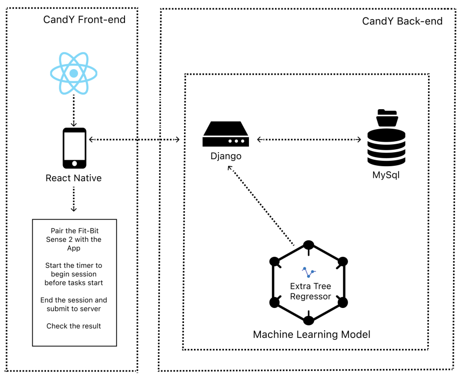

Fig. 4. System Architecture

A. App Front-End

The front-end section is developed by using React Native. This paper used React Native to deliver both an Android and iOS application [35].

There are two ways to program applications utilizing React Native, the Expo platform is selected to develop our application as it allowed for prototyping and real-time evaluation of the data. Expo is an open-source platform for making universal native apps for Android, iOS, and the web with JavaScript and React [36].

B. App Back-End

The Back-end Section is developed by using Django web framework. Django is a high-level Python web framework that encourages rapid development [37]. Also, MySQL is used for the database.

C. Machine Learning Model

The machine learning model section operates the Extremely Randomized Tree that already described above. The model calculates the concentration score by analyzing the user's bio-data. The calculated score is sent to the backend, and the score is transmitted to the front and presented to the user.

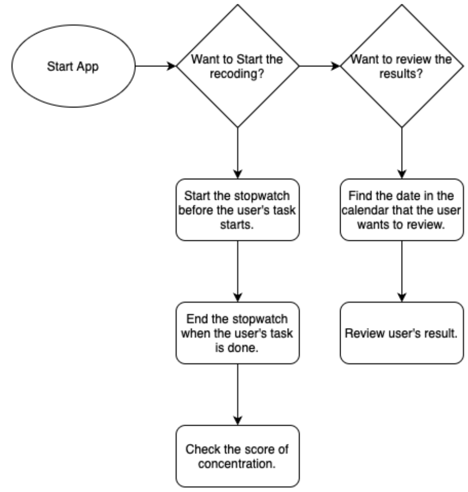

Fig. 5. Flowchart

D. Use-Case

As shown as Fig. 5, two of use-cases exist. With the application developed through React Native and Django, concentration score would be served to users, which is calculated with bio-data, place, and the amount of time. And also, concentration score would be used to optimize their productivity. The primary use-case is as follows.

First, the user should launch the application before starting the work-session.

Second, the user utilizes the stop watch within the application to record their work-session. When the user starts to record their work-session, the app posts the information of work-session to the server. The server save the information to database, MySQL. The database has a table with primary key named user_id and session_id. Once saved, the data can be delivered whenever requested.

Third, the user is provided with a concentration score from the bio-data. The machine learning model calculates the score of the user's concentration with the bio-data features such as HRV, HR, EDA, and Sleep Duration.

VI. Conclusion

This study established the correlation between bio-data and concentration using directly extracted EEG data. Based on this result, the study provides users with meaningful information about concentration. In summary, first, the correlation between EEG and concentration was revealed through the literature reviews. R index value and β/θ SP ratio are the candidates to set the concentration indicator. For these two indicators, Pearson correlation analysis is performed between the value and attention score of EEG sensor, respectively. As a result, β/θ SP ratio showed a stronger relationship than R index and it is adopted as the indicator. Second, the correlation between bio-data and EEG is analyzed. Data processing is performed on the raw bio-data collected by EEG sensor and fitbit device. To analyze the correlation between bio-data and SP ratio, nine regression models are applied. Among them, Extra Tree Regressor(ETR), Random Forest Regressor, XGBoost Regressor showed the highest performance in that order. The optimized ETR model with the best performance is adopted as the final model. Finally, the concentration score is predicted based on the correlation between bio-data and concentration. The system built with React Native front-end and Django Server provides useful information to the user.

REFERENCES

[1] Wilmer, H., Sherman, L. & Chein, J. Smartphones and cognition: A review of research exploring the links between mobile technology habits and cognitive functioning. *Front. Psychol.*. **8** (2017,4)

[2] Noachtar S, Binnie C, Ebersole J, Mauguière F, Sakamoto A, Westmoreland B. A glossary of terms most commonly used by clinical electroencephalographers and proposal for the report form for the EEG findings. *The International Federation of Clinical Neurophysiology. Electroencephalogr Clin Neurophysiol Suppl*. 1999;52:21-41. PMID: 10590974.

[3] Choi, T., Kim, J., Jin, S. & Yoon, G. 8 Determination of the concentrated state using multiple EEG channels. *Electrical, Automation, Control And Information Engineering*. **8**, 1373-1376 (2014)

[4] Liu, N., Chiang, C. & Chu, H. Recognizing the degree of human attention using EEG signals from mobile sensors. *Sensors (Basel)*. **13**, 10273-10286 (2013,8)

[5] Lokare, V. & Netak, L. Concentration level prediction system for the students based on physiological measures using the EEG device. *Intelligent Human Computer Interaction*. pp. 24-33 (2021)

[6] Hasegawa, C. & Oguri, K. The effects of specific musical stimuli on driver s drowsiness. *2006 IEEE Intelligent Transportation Systems Conference*. (2006)

[7] Lutsyuk, N., Éismont, E. & Pavlenko, V. Modulation of attention in healthy children using a course of EEG-feedback sessions. *Neurophysiology*. **38**, 389-395 (2006,9)

[8] Kwon, J. W., Sung, S., Lee, S. B., Lee, H. M., Moon, S. H., & Lee, B. H. (2020). Intraoperative real-time stress in degenerative lumbar spine surgery: simultaneous analysis of electroencephalography signals and heart rate variability: a pilot study. *The spine journal : official journal of the North American Spine Society, 20(8)*, 1203–1210.

[9] Martínez Vásquez, D.A.; Posada-Quintero, H.F.; Rivera Pinzón, D.M. Mutual Information between EDA and EEG in Multiple Cognitive Tasks and Sleep Deprivation Conditions. *Behav. Sci*. 2023, 13, 707.

[10] Kim SC, Lee MH, Jang C, Kwon JW, Park JW. The effect of alpha rhythm sleep on EEG activity and individuals' attention. *J Phys Ther Sci*. 2013 Dec;25(12):1515-8. doi: 10.1589/jpts.25.1515. Epub 2014 Jan 8. PMID: 24409009; PMCID: PMC3885828.

[11] Alba, G., Vila, J., Rey, B., Montoya, P., & Muñoz, M. Á. (2019). The Relationship Between Heart Rate Variability and Electroencephalography Functional Connectivity Variability Is Associated With Cognitive Flexibility. *Frontiers in human neuroscience*, 13, 64.

[12] Aminosharieh Najafi T, Affanni A, Rinaldo R, Zontone P. Driver Attention Assessment Using Physiological Measures from EEG, ECG, and EDA Signals. *Sensors*. 2023; 23(4):2039.

[13] Van Dongen, H. P., Maislin, G., Mullington, J. M., & Dinges, D. F. (2003). The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep*, *26*(2), 117–126.

[14] Benedek, Mathias, and Christian Kaernbach. "A continuous measure of phasic electrodermal activity." *Journal of neuroscience methods 190*.1 (2010): 80-91.

[15] "What should I know about Fitbit sleep stages?", *FitBit*, [Online], Available:https://help.fitbit.com/articles/en_US/Help_article/2163.htm

[16] "How can Fitbit help me track my temperature?", *Fitbit*, [Online], Available:https://help.fitbit.com/articles/en_US/Help_article/2458.htm

[17] "What is Brainwave(EEG)?", *Flowtime Help Center* ,[Online], Available:https://help.meetflowtime.com/Glossary/What-is-Brainwave

[18] "What is Attention?", *Flowtime Help Center* , [Online], Available:https://help.meetflowtime.com/Glossary/What-is-Attention

[19] "What is Heart Rate Variability(HRV)?", *Flowtime Help Center* , [Online], Available:https://help.meetflowtime.com/Glossary/What-is-Heart-Rate-Variability#fnref-1

[20] "What is Heart Rate?", *Flowtime Help Center*,[Online], Available:https://help.meetflowtime.com/Glossary/What-is-Heart-Rate

[21] "Pearson correlation coefficient", *wikipedia*, [Online], Available: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[22] "Linear interpolation", *wikipedia*, [Online], Available: https://en.wikipedia.org/wiki/Linear_interpolation

[23] Peter Blomstrand, Jan Engvall, Effects of a single exercise workout on memory and learning functions in young adults – A systematic review. *Translational Sports Medicine published by John Wiley & Sons Ltd*. 2020;00: 1-13

[24] "Introduction to Data Imputation", *simplilearn*, [Online], Available: https://www.simplilearn.com/data-imputation-article

[25] Li P, Stuart EA, Allison DB. Multiple Imputation: A Flexible Tool for Handling Missing Data. *JAMA*. 2015 Nov 10;314(18):1966-7. doi: 10.1001/jama.2015.15281. PMID: 26547468; PMCID: PMC4638176.

[26] "What is random forest?", *IBM*, [Online], Available: https://www.ibm.com/topics/random-forest

[27] "What is the k-nearest neighbors algorithm?", *IBM*, [Online], Available: https://www.ibm.com/topics/knn

[28] "What is Feature Scaling & Why is it Important in Machine Learning?", *Medium*, [Online], Available: https://towardsdatascience.com/what-is-feature-scaling-why-is-it-important-in-machine-learning-2854ae877048

[29] Södergård, C., & Laakko, T. (2023). Inferring Students' Self-Assessed Concentration Levels in Daily Life Using Biosignal Data From Wearables. *IEEE Access*, 11, 30308-30323.

[30] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[31] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63, 3-42.

[32] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[33] Lundberg, S. M., & Lee, S. I. (2017). *A unified approach to interpreting model predictions. Advances in neural information processing systems*, 30.

[34] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster and U. Ehlert, "Discriminating Stress From Cognitive Load Using a Wearable EDA Device," *in IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 410-417, March 2010, doi: 10.1109/TITB.2009.2036164.

[35] "React Native - Learn once, write anywhere", *React Native*, [Online], Available: https://reactnative.dev

[36] "Expo", *Expo*, [Online], Available: https://expo.dev

[37] "Django: The web framework for perfectionists with deadlines", *django*, [Online], Available: https://www.djangoproject.com/