HMI Project: Predicting Mental Well-Being from Behavioral Engagement Patterns — A Comparison of Machine Learning Models

Kyoheon Hwang [1]
Data Science in Human Behavior, University of Wisconsin-Madison
khwang25@wisc.edu

Abstract

Background: App-based mindfulness meditation programs have become well-established in recent years, leading to numerous studies examining their effects on various psychological symptoms and aspects of mental health. However, the majority of this research has primarily focused on statistical inferential analysis rather than predictive modeling.
Objective: This study first aims to examine user engagement effect on the post-intervention well-being while controlling for baseline mental well-being, perceived stress and prior meditation experience. Subsequently, it also aims to apply machine learning modeling to identify the optimal architecture for predicting mental well-being with the best predictive accuracy.
Method: We first employed Pre-Post ANCOVA to analyze the effect of three key user engagement aspects—Frequency, Consistency, and Track Adherence—and several covariates on mental well-being. For the machine learning phase, the dataset (19,832 observations, 19 features) was partitioned into held-in and held-out sets, followed by multiple imputation on the training data to create 10 imputed datasets. Finally, we fitted and optimized KNN, Random Forest, and Multi-layer Perceptron models across these datasets. The top-performing models were then refitted using the full held-in set and compared against a baseline model to identify the superior architecture.
Results: The inferential model was highly significant ($F(8, 12386) = 1896$, $p < .001$), explaining 55% of the variance in post-intervention well-being ($R^2 = .55$). Specifically, both total app-use days ($b = .023$, $t(12386) = 6.59$, $p < .001$) and LCS score ($b = -0.153$, $t(12386) = 4.00$, $p < .001$) significantly influenced post-intervention well-being. In the machine learning comparison, the MLP model achieved the strongest predictive performance ($R^2 = 0.57$, $RMSE = 1.4721$), marginally outperforming the simple regression baseline ($R^2 = 0.55$) in terms of the coefficient of determination.
Conclusions: This study successfully provided informative prediction of mental well-being scores based on user engagement patterns and diverse demographic variables. For future research, efforts should focus on increasing app retention rates and incorporating other influential features to further enhance model predictive performance.

1. Introduction

Derived from Buddhist and Hindu contemplative traditions, The term "meditation" refers to practices that bring mental capacities and process under greater voluntary (Walsh, 2006), and Mindfulness is a popular style of meditation that is designed to cultivate the non-judgmental awareness in the present moment (Kabat-Zinn, 2005). Meditation and mindfulness-based interventions have been widely developed and applied across clinical and academic fields; however, they have been especially effective in improving quality of life and mental well-being in non-clinical populations. These interventions are known for their soothing effect in reducing the symptoms of anxiety, depression, and stress (Gonzalez-Valero, 2019). With the experience of the COVID-19 era, due to certain features of traditional meditation-such as being conducted in a group setting and requiring a specific physical space-have sometimes acted as barriers to participation for people. In response to these challenges and the rapid development of information technologies, mindfulness-based interventions (MBI) have increasingly been delivered via mobile apps, allowing people to practice meditation without physical restrictions. We have a lot of knowledge on the efficacy of mobile mindfulness applications. Some studies found effects for reducing depression, anxiety and stress (Bostock, 2019) but overall body of literature dealing with app-based MBI is still small (van Emmerik, 2020). Recent study also showed the effect of mobile-based meditation practices on mental health of workers in high-risk professions.

Pace TWW demonstrated the significant reductions in firefighters' anxiety, burnout, and negative affect, from before to after use of the meditation app.

In this context, Healthy Minds Innovations (HMI) provides a valuable platform for large-scale, app-based MBI research. Founded in 2014 by neuroscientist Dr. Richard Davidson, HMI is a non-profit organization dedicated to translating science into tools that cultivate and measure well-being. It developed a free, science based mobile application designed to support mental well-being. The app organizes mental well-being into four trainable pillars (Awareness, Connection, Insight, and Purpose) and offers daily meditation practices and lessons aligned with each pillar. It also features the Healthy Minds Index, a self-report measure of psychological well-being using a 5-point Likert scale. Goldberg SB provides evidence of the HMI app in reducing distress and improving outcomes related to well-being, including social connectedness. Beyond this study, numerous others have also demonstrated the HMI app's effectiveness for mental well-being from an inferential perspective.

Despite a substantial body of inferential research about HMI application effectiveness, non-parametric prediction using machine learning techniques is still rare. Yet such predictive models can analyze users' in-app histories to identify when they are likely to stop using the app, what user characteristics are associated with long-term retention, and the extent to which mental well-being pillars improves. These practical insights can directly inform how the organization iterates and improves the effectiveness of the application and ultimately, this could attract more users to the app.

2. Related work

Goldberg SB executed fully remote RCTs (Randomized Controlled Trials) compared 8 weeks of HMP (Healthy Minds Program) conditions with a waitlist control. In conclusion, this study provides evidence of efficacy for HMP app in reducing distress and improving well-being. Similarly, Schulte-Frankenfeld demonstrated whether mobile-based mindfulness meditation program improved perceived stress, self-regulation and life satisfaction in part-time working university students. He showed the online mindfulness program was found to significantly decrease perceived stress ($\eta_p^2 = .180$, a large effect). It also increased self-regulation ($\eta_p^2 = .195$, a large effect), mindfulness (η_p^2 =.174; a large effect) and cognitive reappraisal ($\eta_p^2 = .136$, a medium effect). In the same way, Cearns M concluded Long-term real-world digital meditation practice is effective and associated with improvements in mood, equanimity, and resilience. These three papers above showed strict effect of mobile meditation on diverse psychological measures including mental well-being. Also, they attempted to use diverse mediators and moderators to their each model, such as days of use, session types, and mood stability. These three papers are highly meaningful in respect to stabilizing positive mental effect of mobile meditation intervention.

Meanwhile, there was other special approach to the mental effect of meditation intervention. Rubin M applied Bayesian Statistic into the study of meditation intervention for posterior prediction. Using Bayesian multilevel models, He found that compared to the waitlist-control, inclusion of mindfulness and compassion intervention led to meaningful reductions in perceived stress b = -3.75, 95% HDI[1] [-6.95, -0.59], anxiety b = -3.79, 95% HDI [-6.99, -0.53], and depression b = -3.01, 95% HDI [-5.22, -0.78]. Rubin's approach on Mindfulness intervention beyond existing studies in the way that by using Bayesian model it could provide probabilistic prediction of posterior distribution that explicitly quantify model uncertainty.

Although various research attempts have been made, the studies mentioned above share a common limitation in a way they were confined to just testing statistical validity of mobile intervention effect rather than model prediction. Despite of Rubin's study that applied a Bayesian model and attempted an interpretation through probabilistic prediction, it still mainly focused on inferential test using Bayesian factor and did not aim to develop more sophisticated predictive models using recent machine learning techniques. Furthermore, all of the aforementioned studies were conducted with relatively small sample sizes of fewer than 400 participants, which limited their capacity to establish a robust predictive model. Finally, although user engagement has been used as an important moderator in several studies, most of them lacked a detailed analysis of it.

2.1. Objective

Thus, in this study, I first examined the effect of user engagement on post-intervention mental well-being using a simple regression framework, with additional analyses to characterize specific engagement patterns. I carefully subdivided user engagement into three aspects: frequency, consistency, and track adherence. This practical approach allows for a detailed understanding of how user engagement contributed to app effectiveness. I then developed predictive models using several machine learning

[1] HDI : High Density Interval

approaches, including a Multi-layer Perceptron (MLP), K-Nearest Neighbors (KNN), and Random Forest. To evaluate their predictive performance, I compared these three candidate models against a baseline regression model to determine whether the machine learning approaches provided improved predictive accuracy. Rather than relying on a single algorithm, I trained and assessed multiple models using a validation set and ultimately selected the one that achieved the highest predictive accuracy. Given that the Healthy Minds Innovations dataset is extensive (nearly two million rows), I was able to secure sufficient data for both training and evaluation.

3. Methods

3.1. Data Collection

When users first log in to the Healthy Minds application, they complete an initial survey that captures demographic information, current perceived stress, and baseline mental well-being(t0). This mental well-being is measured using a set of 5-point-Likert scale questionnaires assessing the four pillars of mental well-being. After four weeks of app use, users complete a follow-up assessment(t1), and a final assessment is administered after eight weeks(t2). During app usage, user activity data such as elapsed time, activity type, activity name and timestamps of usage are recorded. Figure 1 illustrates the data collection procedure from the application in greater detail.

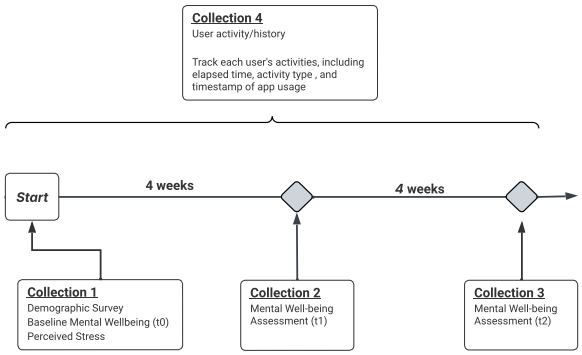

Figure 1. Data Collection Procedure

3.2. Data Cleaning

After concatenating the six datasets, the combined dataset contained 181,356 rows and 26 columns. We then removed duplicate records and excluded users with missing values in essential variables such as mental well-being scores. Finally, we filtered out user activity logs with implausible elapsed times, specifically those shorter than 60 seconds or longer than 4,000 seconds. After executing these data cleaning procedure, we had final dataset contained 19,832 rows and 13 columns. Also, to reduce overfitting in both inferential and predictive model, We diminished dimensionality of categorical variables by merging rear categories. Table 1 provides a summary of the number of categories for each categorical variable.

3.3. Measures

3.3.1. Outcome Variable

The outcome variable was the mental well-being score assessed after four weeks of app use (wellbeing_t1_m). This score represents the average of the four pillars of mental well-being—Awareness, Connection, Insight, and Purpose (ACIP)—as subcategorized by HMI. Each pillar score, assessed using four questionnaire items measured on a 5-point Likert scale, ranges from 0 to 20.

3.3.2. Predictor Variables

Baseline Information

There are a total of eight baseline user information variables, ranging from perceived stress to demographic characteristics, which all users complete after downloading the application. Table 1 presents the data dictionary for these baseline predictors, detailing each variable's name and measurement description.

Table 1: Data Dictionary

Variable Name	Dtype	Description
stress_total	int64	Sum of user's baseline perceived stress, measured by 10 questionnaire items on a 5-point Likert scale.
education	object	Dichotomous variable indicating whether the user holds an advanced degree (beyond college) or not.
age	object	Multi-categorical variable indicating user age with three categories: 18–34, 35–54, and 55 or older.
relationship	object	Dichotomous variable indicating whether the user has partner or not.

Variable Name	Dtype	Description
pre_med_exp	object	A multi-categorical variable capturing users' prior meditation experience, categorized as no prior experience, less than once, 1–2 times, 3–4 times, 5–6 times, and 7 or more times.
gender	object	Multi-categorical variable indicating user gender: Male, Female, and Other
race	object	Dichotomous variable indicating user race whether the user is White or not.
well-being_t0_m	float64	The average score (0 - 20) of the four pillars of mental well-being—Awareness, Connection, Purpose, and Insight. Each pillar was measured using four questionnaire items on a 5-point Likert scale.

User Engagement

User engagement is composed of three key aspects-Frequency, Consistency, and Track Adherence.

Frequency : The frequency represents how often people used the application. I made three new features, which are sum of each participants elapsed time(total_elapsed_time), the total number of days of usage(days_used) and total number of completed sessions(total_session_completed). After checking the correlation matrix and the VIF[2], I left out two problematic variables and used only days_used.

Consistency : The consistency indicates how regularly people use the app. I used the concept of the Coefficient of Variation (CV), which is the standard deviation divided by mean. By doing so, I could quantify how much daily usage time changes from day to day (cv_daily_elapsed_time), and how much the time interval between app-use days varies (cv_interval). A smaller Coefficient of Variation indicates greater consistency in the data.

Track Adherence : The final component of user engagement is track adherence, which reflects how closely users follow the application's recommended meditation track. To quantify this, I used the Longest Common Subsequence (LCS) concept, a measure of sequence similarity that identifies the longest ordered subsequence shared by two sequences. The LCS score is computed as the LCS length divided by the length of the standard sequence, providing an index of how well a user's activity history aligns with the predefined track in the application.(lcs_score_sum).

[2]VIF : Variance Inflation Factor. VIF is a measure used to assess multicollinearity among predictors. A VIF value above 5 typically indicates a potentially problematic level of multicollinearity

4. Experiment

I first examined whether user engagement patterns are really effect on improvement of mental well-being by using Pre-Post ANCOVA approach with simple regression. In addition, I applied machine learning methods including deep learning method to develop predictive models of post-intervention mental well-being and sought to determine the model that achieved the highest level of predictive accuracy.

Figure 7 illustrates the overall workflow of machine learning modeling and model comparison to figure out best performed model. The total dataset, consisting of 19,832 rows and 13 columns, was first split into a held-in set(80%) and a held-out set(20%). The held-in set was then further divided into training set and validation set for model selection. Multiple imputation was performed on the training set using the MICE algorithm, resulting in 10 imputed datasets. Each imputed dataset was fitted using three algorithms-KNN, Random Forest, and Multi-layer Perceptron-to identify the best model with optimized hyper-parameters. After identifying the best-performing KNN, Random Forest, and MLP models using the validation set, we refitted each model on the fully imputed held-in set and computed the RMSE/R^2 using the held-out set. Finally, we compared these metrics with those from baseline model (a basic regression model) to determine which models outperformed the baseline and achieved the highest predictive accuracy.

4.1. Multiple Imputation

Multiple imputation involves filling in the missing values multiple times, creating multiple "complete" datasets. Described in detail by Schafer J.L., the missing values are imputed based on the observed values for a given individual and the relations observed in the data for other participants, assuming the variables are included in the imputation model(Azur MJ, 2011). Multiple imputation by chained equations (MICE) has emerged in the statistical literature as a prominent method to addressing missing values. Creating multiple imputations rather than just single imputation, accounts for the statistical uncertainty in the imputations. Also, this chained equations approach is very flexible and can handle variables of varying types as well as complexities such as bounds or survey skip patterns(Azur MJ, 2011). In this study, I used the *miceforest* package in Python, which provides fast and memory-efficient MICE imputation using random forests. After five MICE iterations without predictive mean matching, I obtained ten complete imputed

datasets.

4.2. Experiment 1 : Pre-Post ANCOVA

Before starting predictive approach using machine learning methods, I tried to verify the effect of mobile based mindfulness intervention using inferential test. I applied Pre-Post ANCOVA to examine user engagement effects on post intervention mental well-being while controlling for perceived stress and previous meditation experience. We used the *statsmodels* package in Python. In this model, user engagement variables- days of app use (days_used), coefficient of variation of daily elapsed time(cv_daily_elapsed_time), coefficient of variation of session intervals(cv_interval), and LCS score (lcs_score_sum)-served as the independent variable, whereas baseline well-being(wellbeing_t0_m), perceived stress, and prior meditation experience were included as covariates.
To include prior meditation experience as a covariate, we regrouped the original categories into three levels (”Frequently”, ”Sometimes”, and ”Never”) and applied planned orthogonal contrast(POC) coding to represent the variable with two contrasts. Contrast 1 compared individuals with any prior meditation experience to those with no experience(”Frequently” : 0.333, ”Sometimes” : 0.333, ”Never” : -0.667). Contrast 2 distinguished between frequent and occasional mediators (”Frequently” : 0,5, ”Sometimes” : -0.5, ”Never” : 0).

4.3. Experiment 2 : Machine Learning Analysis

Before fitting diverse machine learning models, we first conducted One-Hot encoding with categorical variables, and Standard Scaling with continuous variables using *sklearn* package in Python. We did not apply standard scaling to the Random Forest model, as scaling can obscure the original distributions and magnitudes of the features, potentially leading to suboptimal splitting points within the trees. The performance of all models was evaluated using the Root Mean Squared Error(RMSE) and R^2 metrics.
K-Nearest Neighbors (KNN) : K-Nearest Neighbors algorithm is a non-parametric, simple concept algorithm used for both classification and regression tasks in machine learning. The principle behind this algorithm is to find a predefined number of training samples closest in distance to the new point, and predict the label or value from these. The hyperparameter k specifies how many nearest data points are used to estimate the target value in this algorithm. The *KNeighborsRegressor* function in the *sklearn* package was used for KNN model. Specifically, we employed the basic Euclidean distance as the distance metric and applied the ’uniform’ weighting scheme, meaning all neighbors were weighted equally. The number of neighbors(k) was tuned with a range of 1 to 200 using a cross validated Grid Search (cv=5) to identify the optimal k-value for each imputed dataset.
Random Forest (RF) : Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and control overfitting. It uses Bagging[3], where each decision tree is trained on a different bootstrapped sample of the dataset and then these predictions of fitted models called ”base learners” are combined into an aggregated prediction. Using the *RandomForestRegressor* function in *sklearn.ensemble*, we explored variety of hyperparameters of this algorithm. The function to measure the quality of a split(criterion) was set to either ’Gini impurity’ or ’entropy’. The number of trees in ensemble forest (n_estimators) was tuned with (500, 1000, 1500, 2000, 2500) and the number of features to consider when looking for the best split (max_features) was tuned with (”sqrt”, ”log2”). Other hyperparameters were kept at their default values.
Multi-Layer Perceptron (MLP) : MLP is another name of Neural Network with multiple hidden layers. Our model is a feed-forward multi-layer perceptron with an input layer accepting 19 features, followed by two hidden layers considering of 30 and 15 neurons. Each hidden layer used either ReLU, Leaky ReLU or just Linear activation functions, followed by dropout with rates of 0.1, 0.3, or 0.5 and batch normalization. The output layer uses a linear activation to produce the regression target. The network was trained with either the Adam optimizer (weight decay = 1e-4) or SGD optimizer (momentum = 0.9, nesterov enabled, weight decay = 1e-4) within a mean squared error(MSE) loss. The learning rate tuned among 0.0001, 0.001, 0.01. Training proceeded for up to 2000 epochs with no early stopping. Batch size was set to 100, 200 or 500.
Baseline Model : We employed a simple regression model with one-hot encoded and stabilized predictors as a baseline to evaluate the comparative predictive accuracy of the models.

4.4. Experiment 3 : Feature Importance

After figuring out the best predictive model among three, we also tried to demonstrate which features most contributed to the model prediction. We used *permutation importance* function in the sklearn package. Permutation feature importance involves randomly shuffling the values of a single feature and observing the resulting degradation of the model’s score. By breaking the relationship between the feature and the target, we determine how much the model relies on such particular feature.

[3]Bagging : Bootstrap Aggregating

Table 2. Descriptive Statistics for Continuous Variables

Variable	Users, n (%)	Mean (SD)
wellbeing_t0_m	19618 (99.9)	12.39 (2.1)
stress_total	19832 (100)	22.15 (4.8)
days_used	19832 (100)	9.27 (9.0)
cv_daily_elapsed_time	16776 (85.5)	0.52 (0.2)
cv_interval	16011 (80.8)	1.38 (0.56)
lcs_score_sum	19822 (99.9)	0.67 (0.7)

Table 3. Descriptive Statistics for Categorical Variables

Variable	Users (n = 19,832) (%)
Gender	
Male	5649 (28)
Female	9620 (49)
Other	258 (1)
Prefer not to say	224 (1)
Missing value	4081 (20)
Age	
18-34	5141 (26)
35-54	6593 (33)
55-	3791 (19)
Prefer not to say	277 (1)
Missing value	4030 (20)
Race	
White	11108 (56)
Other	3982 (20)
Prefer not to say	624 (3)
Missing value	4118 (20)
Education	
College or higher	13855 (70)
Lower than college	1440 (7)
Prefer not to say	508 (3)
Missing value	4029 (20)
Relationship	
Has partner	10257 (52)
No partner	4991 (25)
Prefer not to say	555 (3)
Missing value	4029 (20)
Meditation Experience	
Has experience	4176 (21)
No experience	11509 (58)
Prefer not to say	116 (1)
Missing value	4031 (20)

5. Result

5.1. Descriptive Statistics

Overall, users had an average baseline mental well-being score of 12.39 (SD = 4.8). On average, users engaged with the application on 9.27 days (SD = 9), exhibited a coefficient of variation of 0.52 for daily elapsed time (SD = 0.2), and showed a coefficient of variation of 1.38 for the interval between app-use days (SD = 0.56). Also, we figured out users average lcs score of 0.67 (SD = 0.7).

In terms of gender, 28% (n = 5649) of users identified as male, 49% (n=9620) as female, 2% as another gender or preferred not to disclose, and 20% had missing data. Regarding age, 26% (n = 5141) were between 18 and 34 years old, 33% (n = 6593) were between 35 and 54, and 19% (n = 3791) were 55 or older. With respect to prior meditation experience, more than half of users (56%, n = 11509) reported no previous experience. See Table 2 and 3 for all descriptive results.

5.2. Pre-Post ANCOVA

The adjusted R^2 for the model was 0.55, and the overall model was significant, F(8, 12386) = 1896, $p < .001$. Regarding user engagement, the total number of app-use days significantly effect on post intervention well-being after controlling for covariates, b=0.023, t(12386) = 6.59, $p < .001$, 95% CI [0.016, 0.030]. In addition, LCS score also showed a significant positive effect on post-intervention well-being, b = 0.153, t(12386) = 4.00, $p < .001$, 95% CI [0.078, 0.229]. This finding suggests that the frequency of app use and users' adherence to the recommended track significantly influence mental well-being, whereas the consistency of app use does not.

In terms of covariates, baseline mental well-being(b = 0.724, t = 107.4, $p < .001$, 95% CI [0.711,0.737]) and perceived stress(b = -0.03, t(12386) = -10.11, $p < .001$, 95% CI[-0.036, -0.024]) had significant effect on post-intervention well-being respectively.

5.3. Machine Learning Analysis

After exploring the predefined hyperparameter ranges for the three machine learning models, we identified the best-performing KNN, Random Forest, and MLP models. We then evaluated the performance of each selected model across the 10 imputed datasets. Table 5 presents the performance metrics of each best-performing model across the imputed datasets, along with their overall averages.

KNN : We found that the best performing KNN model used a k-value of 45, with average RMSE and R^2 values of 1.52 and 0.501, respectively, across all imputed datasets.(Table 5) Figure 2 represents the change of KNN performance across k value (imputation 9). It shows that the R^2 curve increases sharply at the beginning and levels off after approximately k = 10. In contrast, the RMSE curve decreases substantially at lower k values, continues to decline until around k = 45,

Table 4. Linear Model Coef (Outcome : `wellbeing_t1_m`)

Model Summary					
$R^2 = 0.55 \quad R^2_{\text{Adj.}} = 0.55 \quad F = 1896 \quad (p < 0.001)$					
Coefficient Estimates					
Variable	β	SE	t-value	p-value	95% CI
Intercept	4.3326	0.135	32.126	$< .001$	[4.068, 4.60]
wellbeing_t0_m	**0.7239**	**0.007**	**107.38**	$< .001$	**[0.711, 0.737]**
stress_total	**-0.0299**	**0.003**	**-10.11**	$< .001$	**[-0.036, -0.024]**
contrast1	0.1407	0.032	4.334	$< .001$	[0.077, 0.204]
contrast2	0.0986	0.050	1.985	0.047	[0.001, 0.196]
days_used	**0.0231**	**0.004**	**6.592**	$< .001$	**[0.016, 0.030]**
cv_daily_elapsed_time	-0.0058	0.067	-0.086	0.931	[-0.138, 0.126]
cv_interval	0.0180	0.030	0.604	0.546	[-0.040, 0.077]
lcs_score_sum	**0.1535**	**0.038**	**4.001**	$< .001$	**[0.078, 0.229]**

Note. SE = Standard Error. CI = Confidence Interval.
Significant predictors are bolded. $p < .001$ is denoted in bold.

and then shows a slight increase thereafter.

Random Forest : The best performing Random Forest model used a n_estimators of 2,500, max_features of square root, and criterion of gini index, with average RMSE and R^2 values of 1.494 and 0.521 respectively.(Table 5).

MLP: Finally, the best performing Multi-layer Perceptron model consisted of two hidden layers with 50 and 30 units, respectively, using linear activation functions and batch normalization in each layer. The optimal configuration also included a batch size of 200, a learning rate of 0.001, 1,000 training epochs, a dropout rate of 0.2, and the Adam optimizer. Figure 4 illustrates the performance curves across training epochs. The oscillations in both the RMSE and R^2 curves gradually diminished as the number of epochs increased, and the curves finally stabilized around 1,000 epochs.

After identifying the best performing models, we refitted each model using the full held-in datasets and evaluated their performance on the held-out set. We then compared the predictive performance of these models with that of the baseline simple regression model. Multi-Layer Perceptron achieved the strongest performance(RMSE = 1.4721, R^2 = 0.57), outperforming both KNN (RMSE = 1.5443, R^2 = 0.510) and Random Forest (RMSE = 1.5116, R^2 = 0.531). In addition, the MLP model surpassed the simple regression baseline (RMSE = 1.4715, R^2 = 0.55), particularly in terms of R^2(Fig 3).

5.4. Permutation Importance

Figure 5 and 6 illustrate the results of the permutation feature importance analysis. For both RMSE and R^2, baseline well-being (wellbeing_t0_m) showed the strongest and most dominant contribution among all predictors, followed by perceived stress(stress_total), the total days of app use(days_used), lcs score, and male gender.

	KNN		RF		MLP	
Imputation	RMSE	R^2	RMSE	R^2	RMSE	R^2
Imp1	1.5277	0.4993	1.4936	0.5214	1.4973	0.5265
Imp2	1.5222	0.5029	1.4924	0.5222	1.4967	0.5269
Imp3	1.5208	0.5038	1.4937	0.5213	1.4988	0.5255
Imp4	1.5311	0.4970	1.4961	0.5198	1.4990	0.5254
Imp5	1.5230	0.5024	1.4934	0.5215	1.4982	0.5259
Imp6	1.5231	0.5023	1.4908	0.5232	1.4990	0.5254
Imp7	1.5228	0.5025	1.4924	0.5222	1.5000	0.5248
Imp8	1.5240	0.5017	1.4962	0.5197	1.4988	0.5255
Imp9	1.5230	0.5024	1.4982	0.5184	1.4976	0.5263
Imp10	1.5258	0.5005	1.4957	0.5201	1.4989	0.5255
Mean	**1.5244**	**0.5015**	**1.4943**	**0.5210**	**1.4984**	**0.5258**

Table 5. Model performances across imputation

6. Discussion

6.1. Principal Findings

This study sought to advance the scientific understanding of how smartphone-based meditation influences mental well-being by moving beyond traditional statistical inference and incorporating a range of machine learning techniques. To do so, we first re-verified the effects of the app-use pattern of user while controlling for baseline stress and well-being, and also prior meditation experience. We then explored variety of machine learning algorithms and compared their

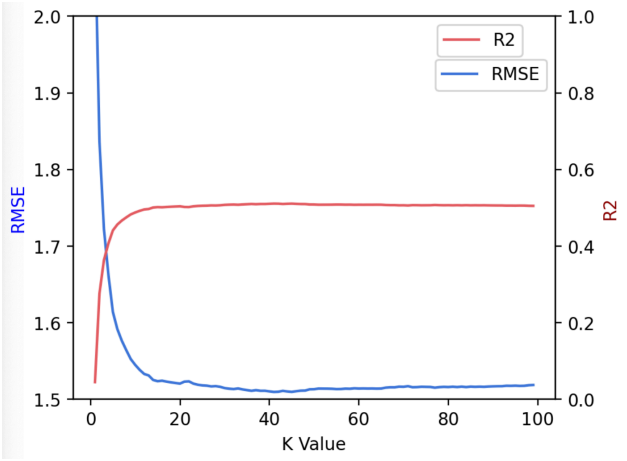

Figure 2. The change of KNN model performance across k-value

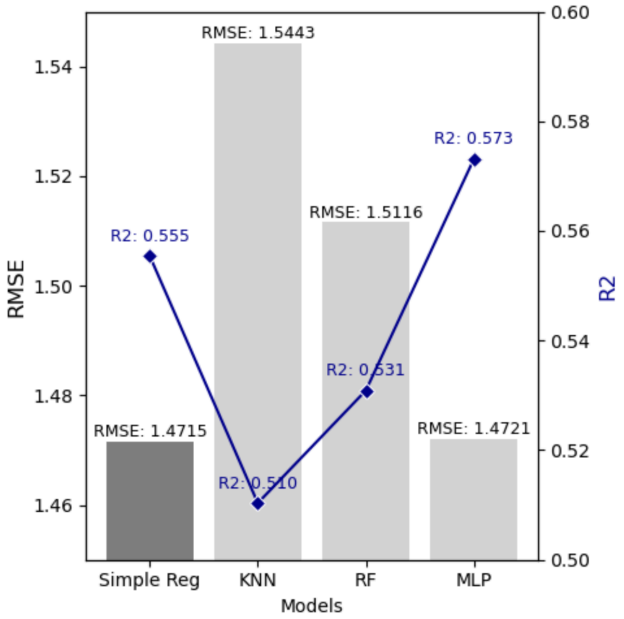

Figure 3. Model Comparison to Selecting Best Model

prediction accuracy to find best predictive performing model. The final Multilayer Perceptron (MLP) model achieved substantially higher predictive power, as measured by the coefficient of determination (R^2), compared to the baseline model. These results suggest that the future mental well-being of HMI app users can be predicted with high accuracy by relying solely on baseline information and short-term usage patterns.

Contrary to out expectations, there was no indication that Consistency of app use produced significant effect on post intervention mental well-being. There are several potential reasons for this. One likely explanation is that consistency truly has no relationship with mental well-being. The other explanation is that, because the four-week interval between pre- and post- intervention assessments represents a relatively short period of app use, it is possible that this timeframe was insufficient to meaningfully capture users' consistency of app engagement. Consequently, the variation in consistency across users may not have contained enough information to explain meaningful differences in mental well-being outcomes.

Regarding the machine learning approach, we found that the Multi-Layer Perceptron achieved the best overall performance and showed a meaningful increase in R^2 compared to the baseline model. However, the optimal activation function for the MLP was a linear function- rather than ReLU or Leaky ReLU-and, somewhat surprisingly, the baseline simple regression model yielded a lower RMSE than the MLP.

This pattern suggests the relationship between the predictors and post-intervention well-being may be predominantly linear, in other words, true Data Generating Process limited the advantage of more complex nonlinear architectures. Figure 5 and Figure 6 support this interpretation by showing that baseline mental well-being was the only predictor that contributed meaningfully to model performance; the remaining predictors, although appearing among the top-ranked features, had only minimal influence. This pattern suggests that, within this dataset, the relationship between baseline and post-intervention mental well-being is primarily linear, with limited additional predictive value provided by user engagement variables, perceived stress, or demographic characteristics. In other words, the dominant role of baseline well-being indicates that users' initial mental-state levels largely determine their post-intervention outcomes, leaving little room for nonlinear or complex interactions to be captured by more flexible machine learning models such as MLP, Random Forest, or KNN.

6.2. Limitation and Future Directions

The notable limitation of this study involves the predictive strength of the variables. Although pre-post ANCOVA analysis confirmed that aspects of user engagement-specifically frequency and track adherence-significantly affected post-intervention mental well-being, their predictive contribution to the model was nominal compared to baseline mental well-being. Furthermore, demographic variables such as gender, race, education, and relationship status offered no meaningful influence on the model's predictions. The inclusion of these categorical variables may have inadvertently led to overfitting; once transformed via one-hot encoding, the resulting increase in binary variables significantly expanded the feature space relative

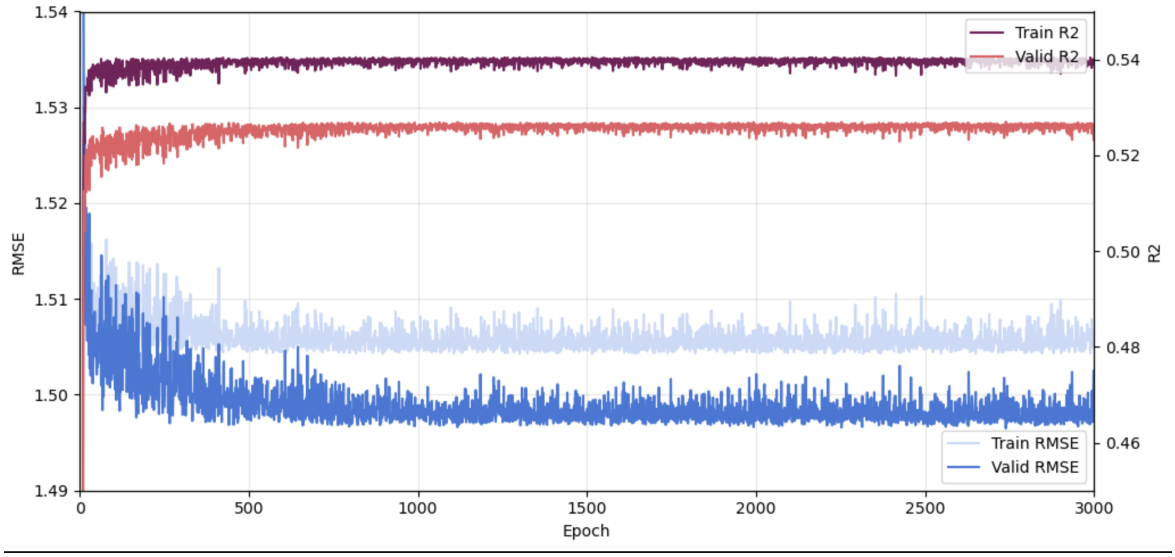

Figure 4. Training and validation performance of the MLP model across epochs

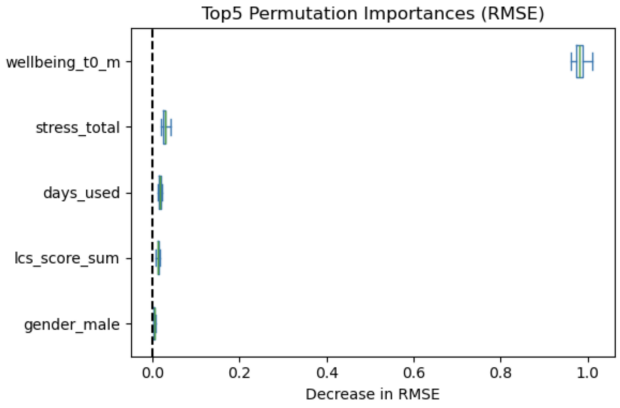

Figure 5. Top5 Permuatation Importances (RMSE)

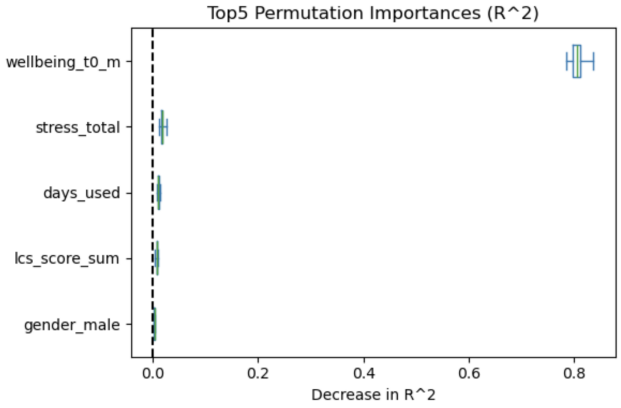

Figure 6. Top5 Permuatation Importances (R^2)

to the sample size.

Furthermore, the best-performing MLP model exclusively utilizing a lineaer activation function (and thus omitting non-linear functions such as ReLU or Leaky ReLU) suggests limitations in both the model's scalability and the utility of the current predictors. This indicates that substantial improvement in predictive performance is unlikely from simply increasing data volume or duration, unless other influential features are identified to challenge the dominant predictive power of baseline well-being. This inference is further verified by the observation that the RMSE of all three best-performing models-KNN, Random Forest, and MLP-was lower than that of a simple regression model (Figure 3).

Consequently, future research is advised to move beyond merely quantifying typical user engagement patterns and prioritize the exploration of other influential features. These features might include the timestamp of user activities or clustered characteristics of users obtained through the Principal Component Analysis (PCA). Furthermore, several completely different avenues should be explored in future work:

- **Alternative Outcomes** : It would be valuable to shift the primary outcome from overall mental well-being to the change in individual perceived stress to better examine its relationship with app-use patterns or mental

well-being scores.

- **Different Models** : Exploring entirely distinct prediction models, such as those focusing on user app retention, could yield significant insights.
- **Granular Analysis** : A more granular investigation should explore how the specific components of mental well-being (e.g., Awareness, Connection, Insight, and Purpose) individually respond to or change based on different app usage patterns.

6.3. Conclusions

Research on mobile-based meditation training has expanded rapidly in the past couple of years, with Healthy Minds Innovation positioned as one of the leading organizations in this field. However, the vast majority of studies associated with this organization have focused exclusively on statistical inferential analysis between clinical/non-clinical psychological metrics and app usage. Our work re-verified that both app use frequency and the level of adherence to assigned tracks are critical factors influencing the app's effectiveness in improving mental well-being. Furthermore, we successfully developed a predictive model that accurately forecasts mental well-being scores, utilizing only a few demographic variables and short-term app usage patterns.

The development of this accurate predictive model offers substantial, actionable value to Healthy Minds Innovation (HMI). Crucially, the model enables HMI to move beyond retrospective inferential analysis to implement a proactive, personalized intervention strategy. By accurately forecasting a user's mental well-being score based on early, short-term usage data, HMI can rapidly identify users who are predicted to have sub-optimal outcomes, thereby streamlining the process of intervention. This predictive capability allows HMI to deploy targeted 'just-in-time' interventions, such as personalized content recommendations or automated nudges, well before significant well-being decline occurs. Ultimately, this shift toward predictive, personalized engagement enhances the scalability of HMI's offerings and maximizes the positive impact of the app on user mental health retention.

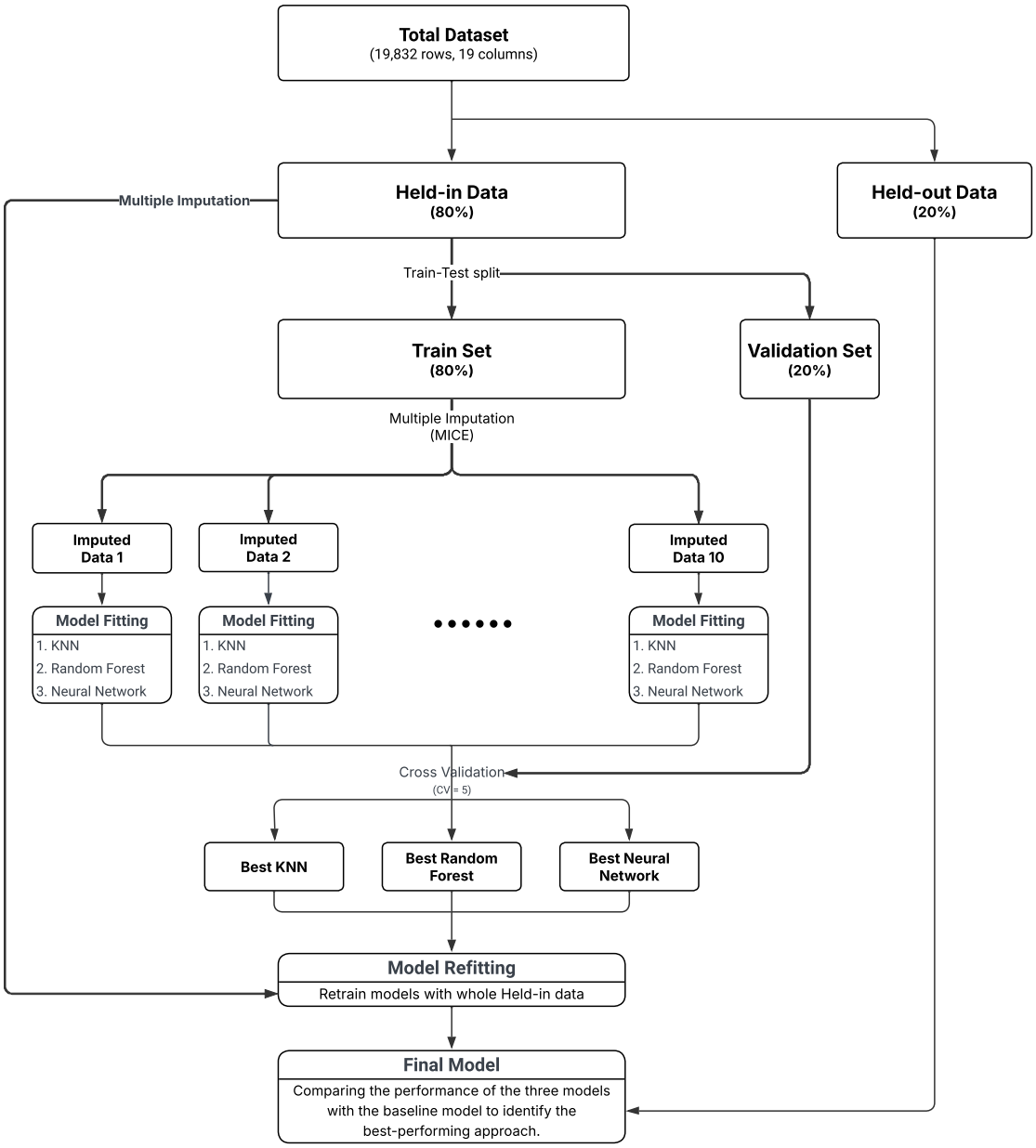

Figure 7. Workflow of overall project

References

Azur MJ, Stuart EA, F. C. L. P. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res*, 20(1), 2011. doi: 10.1002/mpr.329.

Bostock, S., C. A. D. P. A. A.-. S. A. Mindfulness on-the-go: Effects of a mindful- ness meditation app on work stress and well-being. *Journal of Occupational Health Psychology*, 24(1):127–138, 2019. doi: https://doi.org/10.1037/ocp0000118.

Cearns M, C. S. The effects of dose, practice habits, and objects of focus on digital meditation effectiveness and adherence: Longitudinal study of 280,000 digital meditation sessions across 103 countries. *J Med Internet Res*, 2023. doi: 10.2196/43358. URL https://www.jmir.org/2023/1/e43358.

Goldberg SB, Imhoff-Smith T, B. D. W.-M. C. D. C.-D. R. R. M. Testing the efficacy of a multicomponent, self-guided, smartphone-based meditation app: Three-armed randomized controlled trial. *JMIR Ment Health*, 7(11), 2020. doi: 10.2196/23825. URL http://mental.jmir.org/2020/11/e23825/.

Gonzalez-Valero, G., Z.-O. F. U.-J.-J. L. . P.-M. P. Use of meditation and cognitive behavioral therapies for the treatment of stress, depression. *International Journal of Environmental Research and Public Health*, 16(22):4394, 2019. doi: https://doi.org/10.3390/ijerph16224394.

Kabat-Zinn, J. (ed.). *Full catastrophe living: Using the wisdom of your body and mind to face stress, pain, and illness*. Delta Trade Paperback/Bantam Dell, 2005.

Pace TWW, Zeiders KH, C. S. S. E.-H. L. M. N.-W. E. T. R. D. R. Feasibility, acceptability, and preliminary efficacy of an app-based meditation intervention to decrease firefighter psychological distress and burnout: A one-group pilot study. *JMIR Formative Research*, 6(6), 2022. doi: 10.2196/34951.

Rubin M, Fischer CM, T. M. Efficacy of a single session mindfulness based intervention: A randomized clinical trial. *PLoS ONE*, 19(3), 2024. doi: https://doi.org/10.1371/journal.pone.0299300.

Schafer J.L., G. J. Missing data: our view of the state of the art. *Psychological Methods*, 2002. doi: 10.1111/1467-9574.00218.

Schulte-Frankenfeld, P. M., . T. F.-M. App-based mindfulness meditation reduces perceived stress and improves self-regulation in working university students: A randomised controlled trial. *Applied Psychology: Health and Well-Being*, 14(4), 2022. doi: https://doi.org/10.1111/aphw.12328.

van Emmerik, A., K. R. . S.-T. Integrating mindfulness into a routine schedule: The role of mobile-health mindfulness applications. *Nutrition, Fitness, and Mindfulness*, pp. 217–222, 2020.

Walsh, R., . S. S. L. The meeting of meditative disciplines and western psychology: A mutually enriching dialogue. *American Psychologist*, 61(3):227–239, 2006. doi: https://doi.org/10.1037/0003-066X.61.3.227.