Mindfulness (2021) 12:2519–2534
https://doi.org/10.1007/s12671-021-01723-4

ORIGINAL PAPER

The Application of Machine Learning to Online Mindfulness Intervention Data: a Primer and Empirical Example in Compliance Assessment

Damien Lekkas[1,2] · George Price[1,2] · Jason McFadden[1] · Nicholas C. Jacobson[1,3]

Accepted: 4 August 2021 / Published online: 27 August 2021

Abstract

Objectives Machine learning models are a promising, yet underutilized tool within the mindfulness field. Accordingly, this work aimed to provide a practical introduction to key machine learning concepts through an illustrative investigation of the association between at-home mindfulness exercise compliance and stress reduction. To further interrogate the currently inconclusive nature of the compliance-outcome association within the mindfulness literature, the illustrative example leveraged a suite of machine learning techniques to highlight the unique affordances and perspectives of the predictive framework.
Methods Foundational information regarding facets of the machine learning analytical process, including model types, data preprocessing, feature engineering, validation, performance evaluation, and model introspection, was presented. With emphasis on providing details and justifications regarding modeling decisions along the way, the work systematically applied these introduced concepts to a real-world data example. This permitted an opportunity to build, introspect, and derive insight from a model tasked to explore dynamics underlying patient compliance to mindfulness exercises within a web-based delivery setting.
Results The constructed machine learning models suggested a moderate correlation of compliance with post-intervention reliable change in stress ($r = 0.349 \pm 0.018$). Model introspection tools further revealed that a combination of both high consistency and high overall average compliance predicts a trend toward greater reduction in self-reported stress.
Conclusions Results of the illustrative study suggested that compliance, in pattern and absolute magnitude, is a significant contributor to online mindfulness therapy outcomes. Moreover, modeling efforts implicate machine learning as a uniquely beneficial paradigm with which to explore nuanced questions in the mindfulness research space.

Keywords Mindfulness · Applied machine learning · Stress · Online intervention · Tutorial

Mental health disorders, most commonly anxiety and depression, affect over 700 million people worldwide and are associated with social, demographic, and economic hardships (Bartel & Taubman, 1986; Titov et al., 2019). Despite previous attempts to expand access to mental health care, less than half of patients who suffer from mental illness report seeking evidence-based treatment (Titov et al., 2019). Some patients wish to seek treatment for their disorders, but are unable to do so because of financial constraints or the limited availability of psychological services (American Academy of Child and Adolescent Psychiatry Committee on Health Care Access and Economics Task Force on Mental Health, 2009). Others choose not to seek treatment for reasons including social stigma, preferences to self-manage, and a limited awareness of both their mental illness and the potential benefits to therapy (Titov et al., 2019). While the global shortage of therapists and psychiatrists is expected to remain constant, access to technologies (including the Internet and the increasingly ubiquitous smartphone) has increased steadily over the past decade, engendering questions relating to the potential applications of existing digital platforms in

✉ Damien Lekkas
Damien.Lekkas.GR@dartmouth.edu

[1] Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, 46 Centerra Parkway, Suite 300, Lebanon, NH 03766, USA

[2] Quantitative Biomedical Sciences Program, Dartmouth College, Lebanon, USA

[3] Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Lebanon, USA

addressing current barriers to mental health treatment (Tal & Torous, 2017). Current app-based treatments for anxiety and depression have made significant progress in distributing resources and interventions to those in need, regardless of their location, availability, or socioeconomic status (Tal & Torous, 2017). Furthermore, many digital mental health interventions allow users to remain anonymous, which may help alleviate the fear and embarrassment that comes with seeking therapy (Bradford & Rickwood, 2014).

One prominent example of this translation of mental healthcare paradigms into the digital forum comes from efforts in the field of mindfulness. Mindfulness-based stress reduction (MBSR) and its mindfulness-based cognitive therapy (MBCT) derivative are secular therapeutic interventions that use meditation, bodily awareness, and the non-judgmental accepting of the present experience in order to reduce stress, anxiety, depression, and pain (Grossman et al., 2004). Mindfulness interventions aim to restore positive emotions and a calm mental state in patients who suffer from psychosomatic disorders (Zhu et al., 2017). Traditional treatments involve 8-week programs consisting of both weekly group classes led by a certified instructor and at-home mindfulness during which participants employ mindfulness meditation themselves (Huberty et al., 2019). Although current studies suggest that these mindfulness intervention programs are effective at reducing stress and anxiety, these programs can be cost-prohibitive and inaccessible (Mrazek et al., 2019).

Given the advent of digital app-based treatments for stress and anxiety, many studies have analyzed the efficacy of digital mindfulness intervention apps in treating patients with a broad range of psychological disorders, including anxiety, depression, and schizophrenia (Huberty et al., 2019). Mindfulness meditation mobile apps such as Calm and HeadSpace found that participants who used the app daily showed significantly lower levels of stress and significantly higher levels of mindfulness and self-compassion than the control group (Economides et al., 2018; Huberty et al., 2019). Early benchmarking of such smartphone mindfulness apps suggests that they increase patient resilience and have the potential to become effective delivery mediums for mindfulness programs when compared with traditional in-person methods (Mrazek et al., 2019). All told, the current literature reflects a promising future for app-based mindfulness treatments which carry the additional benefits of lower cost and greater accessibility.

In recent years, the efficacy of digital mental health interventions (both app-based and online) has been further probed by leveraging the advantages of machine learning methodologies to consider the dynamic and statistical properties of the temporally collected data that tends to characterize these efforts (Triantafyllidis & Tsanas, 2019). More specifically, machine learning has been used to investigate app-based and online health interventions that target depression management (Burns et al., 2011), stress management (Morrison et al., 2017), and weight loss (Manuvinakurike et al., 2014). Unfortunately, however, despite its broader success in mental health application, little to no research has utilized machine learning to interrogate outcomes, probe dynamics, or assess the efficacy of online-based mindfulness interventions.

In an effort to offer a potential expansion to the analytical toolkit within the mindfulness space, this paper will serve as a practical primer to introduce mindfulness researchers to the paradigm of machine learning and demonstrate, through a real-world data-driven example, the potential utility of machine learning to address more complex questions within the field of mindfulness. With a focus on online-based intervention research, this work will demonstrate how to interrogate the results of a mindfulness intervention study via the construction of a hypothesis-driven machine learning model. The work will thus begin by providing an explanation of machine learning, key components of the modeling architecture and process, as well as the major decisions faced from initial data handling to the interpretation of results. Given the depth and complexities across this broad class of models, the work is not meant to be exhaustive in its treatment, rather it is designed to provide mindfulness researchers with a baseline understanding of modeling construction, choice, assessment, and interpretation to foster and encourage future work of its kind. Accordingly, the remainder of the introduction covers (i) types of models, (ii) pre-processing and missingness of data, (iii) feature engineering, (iv) external validation methods, (v) model selection, (vi) model evaluation, and (vii) model introspection, to provide background and context into the decisions that go into the construction of a machine learning framework as well as information on unique strengths of the approach as compared to more traditional statistical models.

Practical Machine Learning

Machine Learning: a Definition

Machine learning can be broadly defined as a subtype of artificial intelligence that allows for computers to both learn and think on their own (Alzubi et al., 2018).

Types of Models

Machine learning approaches can be broken into three major methods: (i) supervised learning, (ii) unsupervised learning, and (iii) semi-supervised learning. Supervised learning learns the relationship of input data and its corresponding label and can be broken down further into instances where the model seeks to correctly predict either a categorical label (classification) or a continuous numerical value (regression). Unsupervised learning seeks to understand the inherent structure of input data without predefined labels, such as mapping conversation topics in mental health-related online forums (Grant et al., 2018). Semi-supervised learning combines the two aforementioned methods by utilizing models with both labeled and unlabeled data (Zhu & Goldberg, 2009).

Pre-processing and Missingness of Data

Particular consideration must be given to data that is going to serve as input to a machine learning model. First, data types must be interrogated to ensure they are in a format that reflects the research goals or hypotheses. Following data type consideration, it is important to address how individual features may influence the model. To avoid unbalanced influence of features on a model, intra-feature standardization or intra-feature value range transformation is encouraged to ensure consistency in model influence across the features.

Additionally, missing data must be handled prior to implementing a machine learning algorithm. There are three types of missing data to consider: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR refers to instances when the probability of the data being missing is random, and thus, there is no relationship between missing data and any other value in the data. MAR refers to instances where the probability of the data being missing depends on the other collected variables, but is not related to the actual missing value. Lastly, MNAR refers to instances where the characteristics of the missing data do not fall under the MCAR or MAR categories, and modeling of the missing data is required to avoid bias (Kang, 2013). In instances where dropping participants with missing information is inappropriate, one may also replace missing values with an estimate (imputation) to conserve sample size and overall input data information. Imputation can take on many forms depending on the type and degree of missingness in the data. A simple replacement method from a pre-existing data point (single imputation) may suffice, while more complex problems may call for the implementation of simulation models to consider the uncertainty involved in imputing a value (multiple imputation) (Jerez et al., 2010).

Feature Engineering

Following pre-processing and imputation of the raw data, one must consider what information is going to serve as the independent variables (features) of the machine learning model. Transformation of the raw data may be required to better address the underlying problem the predictive model is trying to solve. Methods of feature engineering on the raw data may include calculations of statistical properties (e.g., arithmetic mean of a feature from the original data) (Maxhuni et al., 2016), alterations of distribution (e.g., log transformation), or creation of new features that are derived from a combination of more than one existing feature (e.g., the difference between values of two existing features). Ultimately, feature engineering can be leveraged to enhance both performance and interpretability of a machine learning model.

External Validation Methods

Consideration on how to best leverage a dataset based on the number of participants and the amount of information collected on those participants introduces another unique advantage to machine learning. Unlike traditional statistical methods, which are reliant on the entirety of a study population to inform a particular model, machine learning allows for a training phase and a test phase. The test phase reflects the model's performance on a subset of participants previously unseen by the model, increasing external validity. Considerations for how to split data is generally context dependent. In general, a large study population may allow for a simple separation of the data into a training set and a test set. The training set is then additionally partitioned into a training set and a validation set. The training set allows for model fitting, and the validation set allows for evaluation of model performance. The test set is the only portion of the data suitable for reporting the final model performance metric. However, in instances where a study population may be too small for a traditional train/test split, an additional partitioning technique, *k*-fold cross-validation, can be utilized. *k*-fold cross-validation consists of randomly partitioning the data into *k* groups, where the model is fit on *k*-1 groups and predictions are generated on the last group of data. A particular instance of *k*-fold cross-validation, leave-one-out cross-validation, partitions the data into *k* groups, where *k* reflects the number of subjects (Grimmer et al., 2021).

Model Selection

Model selection is a process that requires consideration of the features utilized in the model, as well as the question being asked of the data. Models can vary based on outcome application (e.g., regression vs. classification), the parameters that control an aspect of the algorithm, known as hyperparameters, and computational efficiency (Grimmer et al., 2021). The segmentation of all machine learning algorithms can be highly heterogeneous; thus, this paper will focus on a select group of model types. *Tree-based* models are a commonly used model type for both classification and regression. Tree-based methods can be broadly broken up into decision trees, random forests, boosting, and Bayesian additive regression trees (BARTs). A tree-based approach divides input data into smaller, uniform groups based on a measure that maximizes separation of the data (Loh, 2011). Tree-based approaches are useful for handling diverse input data, and are computationally efficient (Kern et al., 2019). *Kernel separators*, such as support vector machines (SVMs), are another commonly used model type in both classification and regression problems. SVMs are based on the creation of a line or hyperplane which separates the data into distinct classes (Dreiseitl et al., 2001). *Cluster* models are generally implemented in an unsupervised machine learning framework. Cluster models seek to identify latent topics or groups within a dataset. Partition clustering utilizes methods to partition the data and then assess the similarity between those formed groups, whereas hierarchical clustering combines data points into clusters and repeats this process on the formed clusters (Kassambara, 2017; Saxena et al., 2017). *Artificial neural networks* (ANNs) were inspired by the neurophysiology of information transmission and consist of connections of neurons with weights associated with those connections. While ANNs have been adopted in many areas of research, they are more computationally expensive and difficult to interpret than the aforementioned models (Dreiseitl et al., 2001; Shanmuganathan & Samarasinghe, 2016).

Model Evaluation

Although model selection is important based on the research question being asked and the type of data being used in the model, the decision on what model performance metric to examine and report requires equal consideration. Further, the type of machine learning model influences the evaluation metrics that can be reported. While difficult to provide an exhaustive list of model algorithms and their respective model evaluation metrics, there are a few commonly reported evaluation metrics that should be discussed. In a classification model, area under the receiver operating characteristic curve (AUROC) is commonly reported (Sokolova & Lapalme, 2009). This metric calculates the area under the plot formed by the true-positive rate (sensitivity) against the false-positive rate (1 = specificity). In instances of regression models, the mean squared error (MSE) and the mean absolute error (MAE) is used. The MSE reflects the degree to which the regression line fits the sample data, so in a machine learning framework, the regression line learned from the training set can be applied to the test set to gauge the difference in predicted versus actual outcome scores. The MAE functions similarly, but considers the sum of the absolute difference between actual and predicted values, without consideration for directionality. Lastly, the coefficient of determination (R^2) is commonly reported in regression models, which evaluates how much of the observed variation in the data is explained by the machine learning model (Handelman et al., 2019). An important distinction between the utilization of R^2 and MSE/MAE is that while an R^2 coefficient can be generally interpreted across studies or research disciplines, MSE and MAE are problem specific. For example, an encouraging MSE value in one field of research may reflect a very poor-performing model in another area of research.

Model Introspection

Machine learning allows for model introspection, which can aid in the interpretability of both a model's performance as well as the respective influence of the features on the model's performance. For example, in classification models, feature importance refers to the individual contribution of a given feature on the model classifier. There are a few commonly used methods to evaluate feature importance in a machine learning model. Local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016) offers information on features that were important for classifying a specific observation. A similar method, Shapley Additive Explanations (SHAP), which extends from Shapley values introduced in cooperative game theory, allows for an interpretation of features based on how their respective directionality and magnitude influence the outcome prediction (Lundberg et al., 2019).

Illustrative Study Example

Many factors have been found to contribute to the efficacy of mindfulness programs in reducing symptoms of stress and anxiety. One such factor is patient compliance—the degree to which a patient correctly and consistently follows treatment protocol. This includes attendance at instructor-led sessions, participation in weekly group discussions, adherence to daily mindfulness exercises at home, and/or completion of assigned homework (e.g., journal entries and self-reflections) (de Vibe et al., 2013). Generally, patient compliance is highest during the first few weeks of mindfulness treatment, during which most patients perceive the mindfulness sessions to be easy and immediately gratifying (Lymeus et al., 2019). Once this novelty wears off, however, patient compliance tends to steadily decrease. Many have sought ways of improving patient compliance in mindfulness intervention therapies, with online and app-based delivery methods offering one potential avenue given their absence of in-person components (Yeo et al., 2019).

It is unclear, however, whether patient compliance alone is a reliable predictor of treatment outcome. Some studies have shown that higher levels of patient compliance lead to greater reduction in stress and anxiety (Davis et al., 2007; de Vibe et al., 2013; Quach et al., 2017), while others have found no significant correlation between compliance and patient outcome (Toneatto & Nguyen, 2007). Studies that address the compliance-outcome relationship typically model compliance as a single summative metric (e.g., the average number of completed exercises over the duration of the study) or in terms of some threshold of absolute magnitude for binary treatment (de Vibe et al., 2013; Lengacher et al., 2009). In either case, there is a failure to address trends and trajectories, such as how regularly a subject participates. There is precedent to assume, based on studies that have observed deterioration over time, that compliance is not a set constant, but rather a complex and dynamic facet of therapy engagement that would benefit from more rigorous, multivariable operationalization. Moreover, this observed intra-participant variability in behavior suggests that patterns of compliance may also play an important role in the analysis and interpretation of intervention outcomes. Despite such ubiquitous observations, there is a lack of interrogation in the literature that deals with the temporal patterns of compliance behavior. Notably, this temporal dimension can be operationalized in terms of consistency—how routine a patient is with therapy engagement—to more completely encapsulate the dynamics of a compliance-outcome association. To the authors' knowledge, no study has explored the mindfulness compliance-outcome association in a multimetric framework, nor has there been an analysis that applies notions of consistency to characterize time-based trends in mindfulness exercise engagement.

To provide a practical illustration for the application of machine learning to the field of mindfulness, the current work will utilize data produced from a previous research endeavor that implemented a web-based mindfulness course (Krusche et al., 2012b). The primary goal of the past study was to test feasibility in delivering a mindfulness-based intervention online. Krusche et al. (2012b) reported a significant decrease in subjects' perceived stress as a result of participation in the intervention; however, given the aim of the work, data analysis relied on basic descriptive statistics and, most notably, did not examine how mindfulness exercise practice (herein framed as compliance) affected the reported change in perceived stress. With the dataset publicly available, there is a valuable opportunity to expand their original analysis and interrogate the dynamics and impact of participant compliance on perceived stress outcome using a predictive machine learning framework. Serving to illustrate the analytical process of machine learning-based modeling, this study was guided by the following questions: (i) Can compliance to mindfulness intervention exercises significantly predict reliable change (Jacobson et al., 1984) in perceived stress from pre- to long-term follow-up during a digital treatment? (ii) What are the relative contributions of compliance-based metrics in the prediction of reliable change in perceived stress from pre- to long-term follow-up?

Method

Participants

Participants ($N = 100$, 26% male, 74% female, average age = 48 years) were self-selected for an 8-week preliminary evaluation of an online mindfulness stress reduction course consisting of ten distinct interactive sessions led by two instructors. Treatment consisted of participation in the online mindfulness course and was supplemented by voluntary, unsupervised exercises throughout each week.

Procedures

The data used as the basis for the example study originated from a previous research effort assessing the efficacy of an online mindfulness course (Krusche et al., 2012b) and made publicly available as a supplement (Krusche et al., 2012a). As discussed in the original work, the online intervention combined elements of both MBSR and MBCT into an 8-week course run by the Mental Health Foundation and Wellmind Media and developed under the guidance of leading UK mindfulness instructors. The course had ten interactive, instructor-led sessions where participants were trained in formal meditation techniques, with additional training and practice through website-accessible video guides and exercise assignments. Each week, participants were asked to practice at least one formal exercise using supplied audio and video clips. Exercise durations varied in type and length with frequency of practice quantified via weekly self-report questionnaire. While the course ran for a total of 8 weeks, participants had the option of taking a break and resuming from where they left off upon return. This resulted in the course lasting a minimum of 4 weeks for some individuals.

Measures

Outcome

The type of outcome (dependent variable), whether continuous, ordinal, binary, or otherwise, dictates the class of machine learning model that is appropriate to employ. For the current data, the Perceived Stress Scale (PSS) (Cohen et al., 1983) was utilized as the outcome metric to quantify self-assessed stress levels before, immediately after, and 1 month after completion of the online course. The PSS is the most prevalent and widely accepted measure of personal stress perception and has evidenced good internal consistency reliability (Cohen et al., 1983). It consists of a ten-item survey allowing for a 5-point (0–4) range of response with "4" denoting the highest level of stress perception. The minimum stress score is therefore 0, and the maximum stress score is 40. The PSS is continuous; therefore, a regression-based machine learning model would be the clear choice here; however, note that if there is a theoretically justifiable reason to stratify a continuous outcome by informed cut-points (e.g., $< 20 =$ "low"; $\geq 20 =$ "high"), then a binary or multi-class classification machine learning approach could be leveraged instead.

Self-reported Compliance

Any subset of the raw data that is not the outcome of interest can potentially serve as predictors or features (independent variables) of a machine learning model. Importantly, as will be discussed under "Data Analyses", the variables that comprise the raw data can also serve (in virtually any mathematical combination) as the basis for the creation of additional, derivative features in a machine learning model.

In the study data, for each week of the course, the participants were instructed to practice at least one mindfulness exercise by utilizing the provided video and audio clips. In an effort to stratify individuals based on the frequency of practice outside of the course sessions, participants were asked to review and score the completion/frequency of each of the activities they performed. Compliance in terms of a binary status of completion or as a frequency of exercise practice was therefore dependent on the nature of the exercise as well as the structure of the associated self-report questionnaire item as discussed in detail below. This resulted in a maximum of 12 categorical measures of compliance idiosyncratic in range to each exercise. Exercise compliance was measured in one of four types of scales. The first consisted of a 1–4 scale indicating the amount of practice across the week: (1) every day/time, (2) most days/times, (3) once or twice/few times, and (4) never/not at all. An example prompt is "During week 2, how often have you been practicing Mindful Breathing?" In total, 6 exercises had compliance measured in this format. These include "Body Scan," "Routine Activity," "Mindful Movement," "Mindful Breathing," "Sitting Meditation," and "Chosen Practice." The second scale was very similar to the first and consisted of a 1–4 metric: (1) three times a day, (2) at least once a day, (3) on some days, and (4) not at all. An example prompt is "During week 3, how often have you been practicing the 3 min Breathing Space?" In total, 2 exercises had compliance measured in this format. These include "Breathing Space (Week 3)" and "Breathing Space (Week 4)." The third scale was measured in the range of 1–3 indicating full (1 = yes), partial (2 = sometimes), or no (3 = no) compliance. Only 1 exercise was measured in this manner. "Stress Awareness" had the prompt of "During week 3, your assignment was to be aware of your reactions to stress, without trying to change them. Did you manage to do this?" Lastly, the fourth scale is measured as a binary (1) yes/(2) no. In total, 3 exercises were measured this way—"Mindful Meal," "Event Awareness," and "Activity Awareness." An example prompt is "During week 2, have you been filling out your Event Awareness Journal?" With these scales, higher values denote *lower* compliance. Supplementary Table S1 presents a summary of all self-report questionnaire items along with their associated scales and schedules.

Data Analyses

Developing the Model Outcome

With a clear definition of the research goals and hypotheses, it is common to redefine or restructure the collected data toward a more germane operationalization of the measured outcome. In this example, as noted in "Measures" above, the primary outcome of the mindfulness intervention study was the PSS assessed at three distinct time points throughout the course. As the current endeavor is interested in assessing response to or "success of" the online mindfulness course as a function of compliance, reframing of the empirical outcomes into a useful outcome to predict within a machine learning approach is warranted.

Using the PSS-derived scores from before (PSS_{BF}) and 1 month after (PSS_{OMO}) completion of the course, a difference stress score was calculated to quantify the long-term change in perceived stress as a result of participation in the digital mindfulness intervention. This can be expressed simply as $DIFF = PSS_{OMO} - PSS_{BF}$. In an effort to more robustly capture response to intervention, the reliable change index (RCI) (Jacobson et al., 1984) was also measured and leveraged as the primary outcome metric of interest. The RCI is a psychometric criterion which accounts for the measurement reliability. While it has been shown to be a statistically reliable operationalization of improvement, it is important to note that notions of clinical reliability are not inherent to its formulation. The RCI is a ratio with the numerator the difference score between measures at two time points and the denominator the standard error of measurement of this difference (SE_{DIFF}). From above, the ratio can be expressed as, $DIFF/SE_{DIFF}$. The standard error of the difference itself is related to the standard error of measurement (SEM) as $\sqrt{2 \times (SEM)^2}$, with SEM derived from the product of the standard deviation across all difference scores (σ_{DIFFS}) and the intraclass correlation coefficient (ICC) of the PSS. For the calculations of RCI in this study, an ICC value of 0.66 was specifically selected based on the results of a rigorous and comprehensive empirical study on the generalizability and psychometric properties of the PSS (Miller et al., 2021). Thus, $SEM = \sigma_{DIFFS} \times \sqrt{1 - 0.66}$ toward the final calculation of RCI.

The resulting utility of any predictive machine learning model, whether in terms of practical performance and/or theoretical insight, is tied to how the researcher chooses to define what the model is trying to predict. It is thus important to ensure that the outcome sufficiently represents, in terms of both field-guided theory and statistical rigor, the phenomenology surrounding the research aims. For this illustrative example, the transformation of raw PSS scores into differences scores represents pointed alignment with the study aims, while subsequent modification of these difference scores into measures of reliable change represents an effort to strengthen the predictive model through previous empirical work and statistical refinement. Comparing significance of change rather than absolute difference ultimately allows for the separation of variability due to a change from variability due to measurement error. While not always possible to the same degree as illustrated for this study, it is best practice to strive for a clear and justifiable representation of a predictive model's outcome.

Pre-processing and Missingness of Data

Following selection of an appropriate outcome, a researcher should turn to introspection of the data more broadly. As any study within mindfulness research tends to collect various types of data across a number of subjects over a prolonged period of time (e.g., studies involving participation in mindfulness-based therapies), it is likely that the data will require some "cleaning" or systematic modification to ensure proper compatibility with predictive models. Inconsistent or errant representations of values or variables (e.g., a numeric placeholder for NA or a nominal variable treated as ordinal) should be addressed. Following this, two important checks regarding missingness and inconsistency in variable value ranges must be performed.

In the example data, simple calculations show that there is an overall missingness rate of 2.8% with 10% ($N = 10$) of participants missing at least one data point. This can be considered a low rate of missingness; however, it is rarely best practice to simply drop subjects with missing information. Performing imputation on the data is a popular option; however, depending on the pattern of missingness and imputation strategy, such an approach can bias a model that relies on the imputed data to inform prediction. A simple way to assess the pattern of missingness is to run Little's test for MCAR (Little, 1988) using the *naniar* package in R (Tierney et al., 2021). For the example study data, the model fails to reject the null hypothesis that the data is MCAR ($p = 0.169$), indicating that the use of imputation should have a minimal to no bias on subsequent modeling outcomes. Coupled with a low missingness rate, it is well justified to proceed confidently with imputation and downstream analyses.

Following missingness analysis, the example performs multiple imputation by random forest (Tang & Ishwaran, 2017) to impute the original dataset into ten discrete sample datasets. The process of this algorithm within multiple imputation operates similarly in principle to the tree-based machine learning model of the same name. Briefly, the algorithm operates by first imputing all of the missing data using the mean (or some other basic statistical property), then for each variable that had missing values, it fits a random forest model (a set of decision trees) on the non-missing fraction of the data to ultimately predict the missing fraction. The process repeats a set number of times with each iteration boasting a random forest model that is trained on more representative derivations of the data. Some of its notable strengths when leveraged to impute missing data, compared with default parametric regression–based models for multiple imputation, are that it does not rely on any distributional assumptions of the data, can accommodate non-linear relations and interactions between variables, and can tolerate issues of collinearity among variables (Shah et al., 2014).

The exact method of imputation as well as the number of imputed datasets to derive is left to the discretion of the researcher. It is critical, however, that there is an awareness of the limitations to, and assumptions of, each method when selecting one to employ. In addition, analysis across a larger number of imputed versions of the original data generally results in higher stability and therefore confidence in model results. This is to say that the downstream modeling and analysis procedures must be conducted on each imputed dataset separately with some plan to reconcile and/or combine the findings of each imputed data-driven model after the fact. Accordingly, all subsequent data cleaning and feature engineering for this study was performed separately on each imputed dataset.

Turning to inconsistencies in the range of values across variables, as discussed in "Measures," the ranges for self-report compliance questionnaire items were not consistent across all exercises. Variables that vary in type or range can impact the ability of a machine learning model to properly utilize the data; thus, it is usually safer to ensure that all data is within-variable standardized. For the example, all compliance measures were individually standardized such that each feature had $\mu = 0$ and $\sigma = 1$.

Feature Engineering

The goal of this study is to interrogate the impact of compliance to an online mindfulness intervention on reliable change in perceived stress. Toward this goal, feature engineering can be viewed as a creative endeavor that seeks to derive abstractions of the data that operationalize and/or parse the general phenomenon of interest in different and potentially informative ways. Compliance is the general phenomenon of interest, yet notions of compliance can extend beyond the simple "completion rates" that offer no temporal context or nuanced patterns of behavior. To look at the impact of some more particular aspects of mindfulness compliance on perceived stress outcome within a modeling framework, eight derivative variables (or features) of compliance were created from the raw data:

(i) *Overall average compliance*: It is customary in machine learning modeling to employ features that are statistical summaries of the raw data. To holistically capture the magnitude (or degree) of self-report compliance, this study averaged the normalized compliance scores across all 12 activities. This quantity is akin to a general, overall compliance rate, for example in studies that measure compliance as the number of days a participant engaged in a mindfulness therapy course out of the total number of days of enrollment in the course.

(ii) *Standard deviation of overall compliance*: Similar to (i), the standard deviation of compliance across all 12 activities is a way to broadly quantify compliance; however, this metric is specifically concerned with the general spread of compliance (being the square root of the variance) across the duration of the study rather than the average magnitude. Note that neither (i) or (ii) can address patterns or trends in compliance throughout the study participation—they are both strictly summative.

(iii–vi) *Mean weekly compliance* (for each of weeks 1–4 where data was collected): Feature engineering can involve quantities that result from the compartmentalization of data into meaningful bins. Because the raw data for this study provides self-report compliance across mindfulness exercises that occurred in distinct weeks, parsing compliance by time is one convenient and potentially useful way to design the building blocks with which to assess subject-specific patterns and trends (see features (vii) and (viii)). For this study, the mean weekly compliance for an individual is the average of the normalized self-report compliance scores across the three assigned exercises that define that week. (See Supplementary Table S1 for a complete list of exercises and their respective questionnaire items.)

(vii) *Root mean square of successive differences (RMSSD)* (Von Neumann et al., 1941): The self-report questionnaire for subject compliance with weekly assigned mindfulness activity spanned a total of 12 activities across 4 weeks with 3 activities belonging to a particular week. Using the average of the normalized self-report compliance measures within each week (features (iii)–(vi)), the RMSSD is calculated as $\sqrt{\frac{(c_{w2}-c_{w1})^2+(c_{w3}-c_{w2})^2+(c_{w4}-c_{w3})^2}{3}}$, where c_{wt} is the mean normalized compliance score across activities in week t. This feature is meant to serve as a summative descriptive statistical quantity that is the primary operationalization of compliance consistency. Compliance consistency (or hereafter just "consistency") is an operationalization of how dynamic/variable a subject is in terms of their participation in, and adherence to, assigned weekly mindfulness exercises throughout the study. A subject with higher consistency exhibits lower fluctuations in compliance through time; however, this metric by itself does not speak to the magnitude of this compliance (i.e., two subjects can have identical consistency with vastly different overall magnitudes. See Fig. 1 for an intuitive illustration of this difference.

(viii) *Slope of overall average compliance*: A predictive feature for machine learning can represent an outcome from a statistical model. For this example, operationalization of the trajectory in self-reported compliance over time can be realized through a simple linear model where weekly mean compliance measures for a subject (features (iii)–(vi)) are regressed on time ($t=1, 2, 3, 4$). The resulting slope, m, of this model ($Y \sim mX$, where $X=[1, 2, 3, 4]$ and $Y=[c_{w1}, c_{w2}, c_{w3}, c_{w4}]$) thus represents the overall trend in average compliance across the 4 weeks.

Model Selection and Implementation

An extreme gradient boosted machine learning model (Chen & Guestrin, 2016) (xgbTree) with ten-times repeated tenfold cross-validation on each of the ten imputed datasets was constructed, and hyperparameter was tuned to maximize performance based on variance explained (R^2) using grid search with the *caret* R package to predict PSS RCI from the eight derived features above. Example code in R to run the model on one imputed dataset through the *caret* package is provided for reference in Supplementary Fig. 1. Interested readers are highly encouraged to consult the comprehensive online resource for *caret* written by the creator, Max Kuhn, for further details and examples (Kuhn, 2019). The xgbTree algorithm operates by constructing decision trees (akin to random forest) in a sequential manner where each subsequent tree in the sequence learns from the mistakes of its predecessor and updates the residual errors accordingly. Intuitively, each decision point along a tree can be thought of as a classification of data points based on the value of one independent variable (or predictor) that describes that data point. One key point is that, as with any "boosting" model, the process converts what would at baseline be a set of weak learners into one final strong learner. The decision to utilize this model (over multivariable linear regression or a kernel separator, for example) was based on its often-cited high performance and execution speed in a variety of disparate machine learning tasks and research contexts. As an added benefit, the model can handle missing data by treating it as a unique value to inform prediction. While imputation

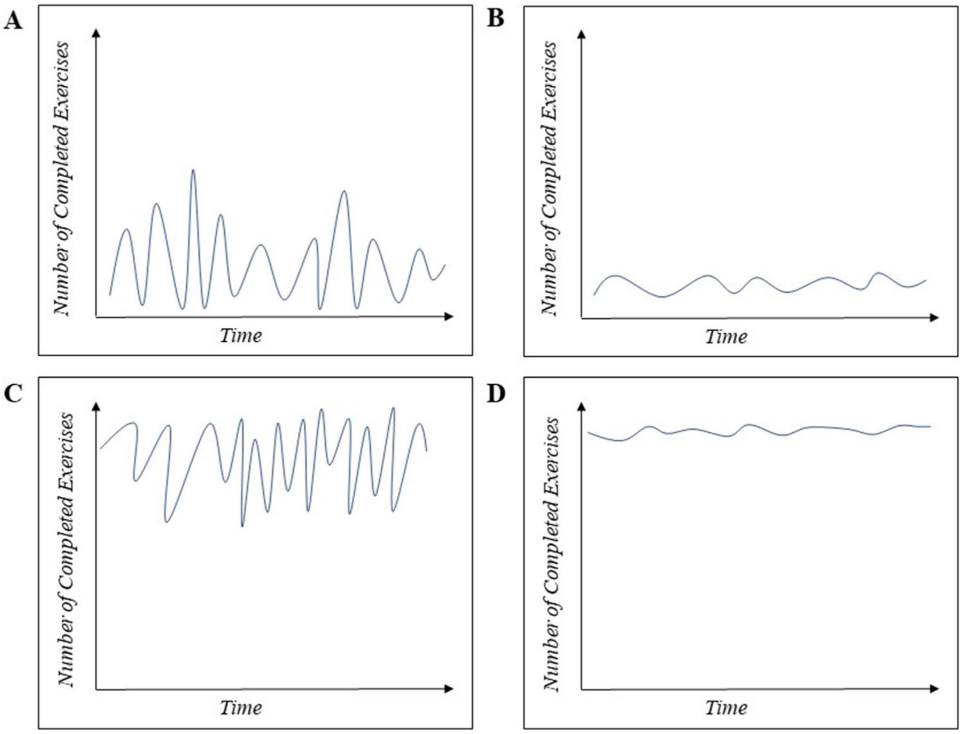

Fig. 1 Overall average compliance versus compliance consistency. These graphs portray hypothetical trends in general compliance over time based on parameters of overall average compliance and consistency in compliance. **A** Low overall average compliance, low consistency of compliance. **B** Low overall average compliance, high consistency. **C** High overall average compliance, low consistency. **D** High overall average compliance, high consistency

was used to handle missing values for this example, in other cases where imputation may not be as appropriate or justified (e.g., high missingness rates, data is not missing at random), extreme gradient boosted models can be employed on the data as-is without the need to resort to subject removal or biasing imputation methodologies.

Model Evaluation

Given that utilized xgbTree models are regression models, performance evaluation and subsequent interpretation for this example is based on the Pearson correlation (*r*) which is the square root of variance explained (R^2).

Relative Feature Importance

Computation of relative feature importance is performed one feature at a time where MAE is measured before and after shuffling the values associated with the feature. This disrupts the association between the outcome and the feature, thus allowing for a comparison of prediction error before and after perturbation. Larger increases in error subsequent to shuffling indicate higher importance. Each feature is compared based on this magnitude of increase to arrive at a relative ranking of importance. The comparative metric for a feature thus indicates the factor by which the model's prediction error increases when that feature's values are modified.

Model Introspection with LIME

The LIME is a prominent model-agnostic explanatory algorithm in machine learning that randomly perturbs the values of the predictive features to appreciate how model prediction changes in response (Ribeiro et al., 2016). When performed iteratively across all data points, and in conjunction with appropriate visualization tools, the resulting prediction values can be utilized to explain "localized surfaces" of the model prediction landscape. As derived and illustrated in Friedman (1999), a partial dependence plot (PDP) is one useful way of visualizing this "black box" introspection of machine learning model prediction. More specifically, it allows for an intuitive presentation of the individual or marginal effect of one or two simultaneous features on the predicted value of the outcome (Friedman, 1999). In this manner, PDPs serve to more thoroughly interrogate aspects of mindfulness exercise compliance and how they are expected to influence stress outcomes.

For the example study, the *iml* R package (Molnar et al., 2018) was used to calculate (i) feature correlations, (ii) relative feature importance, and (iii) partial dependence along with associated graphs for each of the eight model features across each of ten imputed datasets. The authors recommend that interested readers consult the comprehensive online guide on *iml* written by Molnar (2021) for more detailed coverage and examples of implementation. Feature importance was averaged across all imputed datasets to arrive at a final representative ranking of importance. Partial dependence calculations were combined and smoothed across all ten imputed datasets using generative additive models (GAMs).

Results

Compliance as a Predictor of Long-term Stress Reduction

The average model performance across the ten imputed datasets for RCI of stress outcome reflects a moderate correlation between compliance and post-intervention stress reduction ($r = 0.349 \pm 0.018$). The results indicate that compliance to an online mindfulness intervention alone accounts for approximately 12.2% of the variance related to long-term stress reduction outcome.

Relative Feature Importance

Among the eight features used to predict RCI of stress, (i) overall average compliance, (ii) slope of overall average compliance, and (iii) RMSSD were the top three ranked features driving predictions of the machine learning model. See Supplementary Table S2 for the complete ranking and importance scores. Ranking scores indicate that perturbation of non-temporal, overall average compliance resulted in a 4.90 factor change in model prediction error, with the temporal features of slope and RMSSD resulting in a 3.99 and 3.65 factor change, respectively.

Partial Dependence and Feature Interaction

The marginal effects of each of the top three most important model features on reliable change prediction suggest that a trend toward higher overall average compliance (Fig. 2A) and higher consistency in compliance, i.e., lower RMSSD (Fig. 2B), as well as greater increases in compliance over time (i.e., slope of mean compliance) (Fig. 2C) are individually/marginally associated with greater reliable change in perceived stress across 8–12 weeks. The joint marginal effect (interaction) of overall average compliance and RMSSD on model prediction illustrates that a trend toward higher levels of overall average compliance in conjunction with a trend toward lower RMSSD (higher compliance consistency) is predictive of a greater reduction in long-term stress (Fig. 2D). The worst outcomes (lowest predicted RCIs) are predicted to occur within ranges of lowest consistency (highest RMSSD) and lowest overall average compliance. By extension, this suggests that being inconsistently, highly compliant (Fig. 1C) with mindfulness exercises is associated with more favorable outcomes than being inconsistently, lowly compliant (Fig. 1A). Furthermore, the contour lines of Fig. 2D highlight that the influence of consistency on RCI is more incrementally impactful across ranges of lower overall average compliance. Taken together, the results indicate that both overall average compliance and consistency are associated with reliable change in self-reported stress, with consistency having a differential impact on predictive dynamics that is modulated by the overall degree of compliance exhibited throughout the course of the therapy.

Discussion

Context and Significance

The study utilized a publicly available dataset with self-reported participant compliance to an online web-based mindfulness intervention to interrogate the impact of mindfulness exercise compliance across 4 weeks on self-reported stress outcome 1 month out. Capitalizing on the advantages of a feature engineering-based machine learning approach, novel measures of compliance were constructed from the data to predict reliable change in PSS. The selected example served as an extension of the research conducted by Krusche et al. (2012b), where the focus was on evaluating the overall efficacy of a fully web-based mindfulness intervention course. To this end, the primary metrics of analysis were changes in mean PSS before and after the course, and after 1 month of follow-up. Statistics were descriptive and reflected a significance in change from pre- to post-intervention ($p < 0.001$) with large effect size ($d = 1.57$) greater than reported by most in-person mindfulness intervention courses at the time (Krusche et al., 2012b). To assess the impact of mindfulness practice on PSS outcome, the cohort was stratified into two and three levels of relative frequency of self-reported practice. They found no significant difference in PSS score between groups with either stratification regime, which was described as a ramification of a potentially too broad and ill-fitting self-reporting template.

The longitudinal nature of self-reported compliance information in the selected dataset was, in fact, a notable strength as it permitted the consideration of temporal trends in mindfulness practice. Instead of quantifying at-home practice in terms of absolute frequency across the intervention period akin to how Krusche and colleagues (2012a, b) stratified their participants, the data for the example was partitioned on a per-week basis to capture more nuanced dynamics of course engagement. A hallmark of the machine learning paradigm is the process of feature engineering allowed for temporal operationalizations of what is normally a static treatment of compliance within the mindfulness literature. Chief among the derived variables was RMSSD, which allowed for a summative metric of change over time to address notions of compliance consistency. While Krusche et al. (2012a, 2012b) succeeded in testing the feasibility of

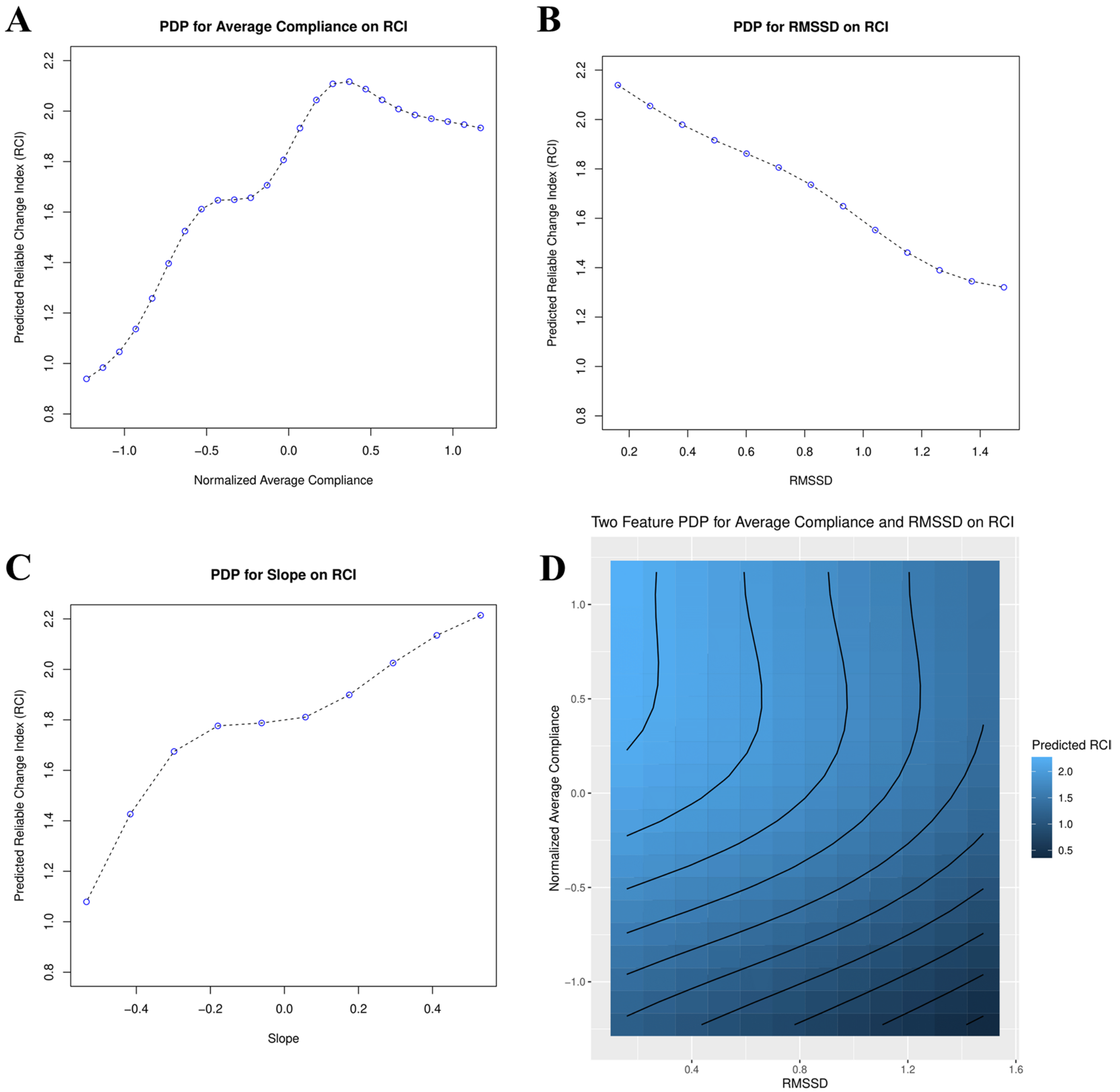

Fig. 2 Partial dependence plots for the top three important predictors driving model prediction. **A** Plot of the marginal effect of overall average compliance on model prediction of reliable change in stress. **B** Plot of the marginal effect of RMSSD on model prediction of reliable change in stress. **C** Plot of the marginal effect of the slope of overall average compliance on model prediction of reliable change in stress. **D** Plot of the joint marginal effect of both overall average compliance and RMSSD on model prediction of reliable change in stress. All reliable change indices are based on reliable change in PSS scores from pre- to long-term follow-up

mindfulness intervention in an online medium with promising results, they did not fully explore the potential impact of self-reported informal practice (compliance) on the outcomes they observed. To demonstrate a practical and useful application of machine learning to the mindfulness body of research, the current study focused on presenting how the relationship between participant compliance and mindfulness intervention outcome can be explored. With a greater proportion of the mindfulness intervention literature reflecting on the accuracy and/or reliability of the compliance measures utilized rather than assessing direct associations on intervention outcome, this example highlighted how, with a proper dataset, predictive modeling can provide the opportunity to contribute to a relatively small and currently inconclusive body of research (Carmody & Baer, 2008; Crane et al., 2014; Hawley et al., 2014; Parsons et al., 2017). Moreover, where previous literature in this space uniformly treats compliance as a summative, static variable, the study leveraged the temporal context of a prospective cohort dataset to expand the modeling conceptualization and consideration of compliance impact on the intervention outcomes observed.

The Machine Learning Advantage

A machine learning predictive model has characteristics that make it an attractive extension of traditional statistical modeling. As distinct from investigations that look at the direct associations of compliance on intervention outcome using descriptive statistics and/or traditional regression techniques and mediation analysis (Bondolfi et al., 2010; Carmody & Baer, 2008; Crane et al., 2014; Fuhr et al., 2018; Hawley et al., 2014; Luberto et al., 2018; Perich et al., 2013; Quach et al., 2017), machine learning methods are capable of assessing the more complex non-linear and interactive dynamics of several derivative compliance variables simultaneously. Additionally, a machine learning model possesses some degree of inherent generalizability, even when the available data is comparatively small (i.e., $N=100$ as in the illustrative example). Indeed, the cross-validation paradigm of machine learning ensures that all results reflect modeling on data that has never been seen while training. Lastly, the implementation of a model-agnostic explainer (i.e., LIME or SHAP) affords a more specific understanding of the relative component contributions that drive overall correlative associations of variables on an outcome. In this way, the behavior of a model (the predictions made) in response to variable perturbation can serve as an indicator of absolute and relative performance of predictors, as well as a means by which to observe the joint effects of variables on the outcome. For the study, this equated to assessment of feature engineered compliance metrics on predicted reliable change PSS outcome. Taken together, a machine learning analytical pipeline of the kind presented in the study results in a very different means of quantifying the explanatory power of compliance on symptom change in response to intervention, which ultimately translated into a clear quantification of compliance impact both holistically (e.g., overall average compliance) and in a temporally parsed reductionist framework (e.g., RMSSD, slope of overall average compliance).

Model Performance Results: Interpretation

The primary result of the machine learning model concerns predictive performance and suggests that compliance to informal exercises assigned as part of an online mindfulness intervention broadly explains 12.2% of the variance in reliable change of PSS from before intervention to 1 month after completion of the course. This equates to a moderate correlation of 0.349 between the independent variables (engineered features) included in the model and reliable change in PSS—the model's dependent variable (outcome to predict). Assessment of what is "acceptable" or "functionally useful" predictive performance for regression models is usually dictated by the details of the predictive task (e.g., phenomenological complexity of the outcome, specific hypotheses/aims of the research, theory-guided expectations from predictive features), as well as the general expectations and benchmarks within the pertinent field/subfield. It may be that a model that is capable of explaining 5% of the variance ($r=0.224$) is useful in one setting, while a model that is capable of explaining 20% of the variance ($r=0.447$) is of no merit in another. Please see Supplementary Discussion 1 for interpretation and commentary of an alternate outcome.

Model Introspection-Feature Importance: Interpretation

Although the model result quantifies the predicted impact of compliance on reliable stress change holistically ($r=0.349$), it cannot reflect the relative effects of each compliance predictor on outcome. By using LIME, however, the illustrative study explored how the model "learned" the data through predictor perturbation. The marginal effects of each predictor on model prediction that result from this methodology provided a structured way of comparing compliance metrics, thereby determining relative contributions of compliance predictors on self-reported mindfulness intervention stress reduction. Each predictor was designed to capture a different aspect of compliance (see the "Feature Engineering" subsection in "Data Analyses"); thus, predictors with significantly higher importance relative to others may speak to dominant roles in certain phenomenological nuances of compliance behavior. The illustrative example results indicated that overall average compliance is most impactful for the model's predictions, substantiating the utility of temporally insensitive measures in the literature; however, measures of compliance trajectory and pattern (i.e., RMSSD and slope) were also found to have an appreciable impact (3.65 to 3.99 factor increase in model error when perturbed; see Supplementary Table S2). This hints that such temporal considerations of compliance (e.g., consistency) may offer additional insight beyond what is possible from more simple summative metrics, and advocates for a complementary application of both time-sensitive and time-insensitive measurements of compliance. Please see Supplementary Discussion 2 for interpretation and commentary of an alternate outcome.

Model Introspection-Feature Interaction: Interpretation

Where feature importance metrics can introspect the relative impact of a feature on the model, they cannot reflect directionality or trends in association between feature and outcome. Therefore, upon finding that features have a substantial and significant relative impact on the prediction decisions of the machine learning model, their single and joint effects on the outcome across a range of values can be probed further in LIME with additional visualization techniques. The partial dependence results graphed in Fig. 2A–C illustrate general trends that are expected: (i) individuals with low average compliance across weeks are predicted to have a lower change in stress reduction compared with individuals with high average compliance across weeks (Fig. 2A), (ii) individuals who are less consistent in their compliance (regardless of whether it is high or low) likewise tend toward lower change in stress reduction compared with those who are more consistent (Fig. 2B), and (iii) individuals who showed a negative trajectory (slope) in average compliance through the study reflected worse outcomes (Fig. 2C). The trend is most apparent and consistent in Fig. 2B with a continuous and gradual decrease in predicted reliable change in stress with increasing RMSSD (decreased consistency).

Further investigation of the joint effect of overall average compliance and compliance consistency (i.e., RMSSD) on outcome suggests that higher consistency, coupled with higher overall average compliance, is predicted to result in more favorable reliable change (Fig. 2D). Leverage of this visualization highlights the utility of investigating the interactive effects of features. For the current empirical example, this translates to an ability to illustrate the synergistic importance of both traditional mindfulness measures of compliance (i.e., overall average compliance) and novel, derived, temporal operationalizations (i.e., RMSSD). See Supplementary Discussion 3 and Supplementary Fig. S2 for interpretation and commentary of an alternate outcome.

Limitations and Future Research

The presented study was a practical example of applying a machine learning methodology to mindfulness research for the extraction of novel insight. To this end, the results evidence potential benefit to extending the manner by which compliance is conceptualized and operationalized. The selected dataset fit well within the broader context of a primer and general introduction to machine learning. As any real-world dataset has its idiosyncrasies which naturally are consequential to the relative pertinence of modeling options and choices, this work was limited in its ability to cover optional methods of, and components to, the machine learning modeling process such as feature autoencoding, manual hyperparameter tuning, pre-processing of text-based data, and oversampling/undersampling techniques. While the goals of this work did not align with exhaustive treatment of the machine learning field, it is nonetheless important for readers to be aware that modeling versatility, extensibility, and customizability extend far beyond what has been presented.

Concerning the present study specifically, the results suggest that future exploration into the impact of compliance on mindfulness intervention outcome is warranted. By extension, consideration of compliance consistency and other metrics that look at the temporal dynamics of compliance throughout the duration of the intervention may be fruitful toward more complete assessments of intervention efficacy. With the ubiquity and continued growth of digital health initiatives and online mental health treatment resources, it is unsurprising that the preferred format to receive mindfulness meditation interventions is online (Wahbeh et al., 2014). Compliance, especially in this individualized and private format, is more complicated to measure. As researchers continue to develop and test new interventions for this digitized translation, analyzing the impact of compliance may require the leverage of more intricate models to offset limitations in its quantification. Machine learning is one avenue with which to mitigate deficiencies inherent to the "unreliable estimates of participants' true informal engagement" that exist across studies (Carmody & Baer, 2008; Crane et al., 2014; Hawley et al., 2014). With no standardized unit of measurement that captures the amount of time, frequency, or degree of "quality" of exercise practice (Segal et al., 2019), the development of many derivative compliance-based predictors within a machine learning framework can more completely capture and analyze compliance effects and can therefore transcend the operationalization of compliance as a single metric. The authors encourage mindfulness researchers to broadly leverage machine learning predictive modeling as a complementary analytical technique in future tests of hypotheses, evaluations of therapies, and explorations of their data.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12671-021-01723-4.

Author Contribution DL: designed and executed the study, built the model pipeline, performed the data analysis and interpretation, created the figures, conducted the literature review, and wrote the paper. GP: conducted the literature review and wrote the paper. JM: conducted the literature review and assisted in writing the introduction. NCJ: assisted in the data interpretation, method refinement, and model development, and assisted in the editing of the final manuscript. All authors approved the final version of the manuscript for submission.

Funding This work was funded by an institutional grant from the National Institute on Drug Abuse (NIDA-5P30DA02992610).

Declarations

Ethics Approval and Consent to Participate No formal ethics approval was needed. Original confirmation was obtained from the Research and Development Ethics Clinical Audit and Effectiveness Department, Oxford. The data utilized were originally taken as a survey obtained by an opt-in system where "implied consent" is relevant. Patient consent was taken upon registering for the course.

Conflict of Interest The authors declare no competing interests.

References

Academyof Child and Adolescent Psychiatry Committee on Health Care Access and Economics Task Force on Mental Health., A. (2009). Improving mental health services in primary care: Reducing administrative and financial barriers to access and collaboration. *Pediatrics, 123*(4), 1248–1251. https://doi.org/10.1542/peds.2009-0048

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: An overview. *Journal of Physics: Conference Series, 1142*, 012012. https://doi.org/10.1088/1742-6596/1142/1/012012

Bartel, A., & Taubman, P. (1986). Some economic and demographic consequences of mental illness. *Journal of Labor Economics, 4*(2), 243–256. https://doi.org/10.1086/298102

Bondolfi, G., Jermann, F., der Linden, M. V., Gex-Fabry, M., Bizzini, L., Rouget, B. W., Myers-Arrazola, L., Gonzalez, C., Segal, Z., Aubry, J.-M., & Bertschy, G. (2010). Depression relapse prophylaxis with mindfulness-based cognitive therapy: Replication and extension in the Swiss health care system. *Journal of Affective Disorders, 122*(3), 224–231. https://doi.org/10.1016/j.jad.2009.07.007

Bradford, S., & Rickwood, D. (2014). Adolescent's preferred modes of delivery for mental health services. *Child and Adolescent Mental Health, 19*(1), 39–45. https://doi.org/10.1111/camh.12002

Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., & Mohr, D. C. (2011). Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research, 13*(3), e55. https://doi.org/10.2196/jmir.1838

Carmody, J., & Baer, R. A. (2008). Relationships between mindfulness practice and levels of mindfulness, medical and psychological symptoms and well-being in a mindfulness-based stress reduction program. *Journal of Behavioral Medicine, 31*(1), 23–33. https://doi.org/10.1007/s10865-007-9130-7

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*. https://doi.org/10.2307/2136404

Crane, C., Crane, R. S., Eames, C., Fennell, M. J. V., Silverton, S., Williams, J. M. G., & Barnhofer, T. (2014). The effects of amount of home meditation practice in mindfulness based cognitive therapy on hazard of relapse to depression in the staying well after depression trial. *Behaviour Research and Therapy, 63*, 17–24. https://doi.org/10.1016/j.brat.2014.08.015

Davis, J. M., Fleming, M. F., Bonus, K. A., & Baker, T. B. (2007). A pilot study on mindfulness based stress reduction for smokers. *BMC Complementary and Alternative Medicine, 7*(1), 2. https://doi.org/10.1186/1472-6882-7-2

de Vibe, M., Solhaug, I., Tyssen, R., Friborg, O., Rosenvinge, J. H., Sørlie, T., & Bjørndal, A. (2013). Mindfulness training for stress management: A randomised controlled study of medical and psychology students. *BMC Medical Education, 13*(1), 107. https://doi.org/10.1186/1472-6920-13-107

Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., & Binder, M. (2001). A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *Journal of Biomedical Informatics, 34*(1), 28–36. https://doi.org/10.1006/jbin.2001.1004

Economides, M., Martman, J., Bell, M. J., & Sanderson, B. (2018). Improvements in stress, affect, and irritability following brief use of a mindfulness-based smartphone app: A randomized controlled trial. *Mindfulness, 9*(5), 1584–1593. https://doi.org/10.1007/s12671-018-0905-4

Friedman, J. H. (1999). *Greedy function approximation: A gradient boosted machine*. Retrieved from https://statweb.stanford.edu/~jhf/ftp/trebst.pdf. Accessed 6 Jul 2020.

Fuhr, K., Schröder, J., Berger, T., Moritz, S., Meyer, B., Lutz, W., Hohagen, F., Hautzinger, M., & Klein, J. P. (2018). The association between adherence and outcome in an internet intervention for depression. *Journal of Affective Disorders, 229*, 443–449. https://doi.org/10.1016/j.jad.2017.12.028

Grant, R. N., Kucher, D., León, A. M., Gemmell, J. F., Raicu, D. S., & Fodeh, S. J. (2018). Automatic extraction of informal topics from online suicidal ideation. *BMC Bioinformatics, 19*(S8), 211. https://doi.org/10.1186/s12859-018-2197-z

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science, 24*(1), 395–419. https://doi.org/10.1146/annurev-polisci-053119-015921

Grossman, P., Niemann, L., Schmidt, S., & Walach, H. (2004). Mindfulness-based stress reduction and health benefits. *Journal of Psychosomatic Research, 57*(1), 35–43. https://doi.org/10.1016/S0022-3999(03)00573-7

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Lee, M. J., & Asadi, H. (2019). Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. *American Journal of Roentgenology, 212*(1), 38–43. https://doi.org/10.2214/AJR.18.20224

Hawley, L. L., Schwartz, D., Bieling, P. J., Irving, J., Corcoran, K., Farb, N. A. S., Anderson, A. K., & Segal, Z. V. (2014). Mindfulness practice, rumination and clinical outcome in mindfulness-based treatment. *Cognitive Therapy and Research, 38*(1), 1–9. https://doi.org/10.1007/s10608-013-9586-4

Huberty, J., Green, J., Glissmann, C., Larkey, L., Puzia, M., & Lee, C. (2019). Efficacy of the mindfulness meditation mobile app "Calm" to reduce stress among college students: Randomized controlled trial. *JMIR MHealth and UHealth, 7*(6), e14273. https://doi.org/10.2196/14273

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15*(4), 336–352. https://doi.org/10.1016/S0005-7894(84)80002-7

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine, 50*(2), 105–115. https://doi.org/10.1016/j.artmed.2010.05.002

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology, 64*(5), 402. https://doi.org/10.4097/kjae.2013.64.5.402

Kassambara, A. (2017). *A practical guide to cluster analysis in R: Unsupervised machine learning*. CreateSpace Independent Publishing Platform.

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods, 13*(1), 73–93.

Krusche, A., Cyhlarova, E., King, S., & Williams, J. M. G. (2012a). *Data from: Mindfulness online: A preliminary evaluation of the feasibility of a web-based mindfulness course and the impact on stress*. [Data guide and codebook]. https://doi.org/10.5061/dryad.f4688

Krusche, A., Cyhlarova, E., King, S., & Williams, J. M. G. (2012b). Mindfulness online: A preliminary evaluation of the feasibility of a web-based mindfulness course and the impact on stress. *British Medical Journal Open, 2*(3), e000803. https://doi.org/10.1136/bmjopen-2011-000803

Kuhn, M. (2019). *The caret package*. Retrieved from https://topepo.github.io/caret/index.html. Accessed 27 May 2021.

Lengacher, C. A., Johnson-Mallard, V., Post-White, J., Moscoso, M. S., Jacobsen, P. B., Klein, T. W., Widen, R. H., Fitzgerald, S. G., Shelton, M. M., Barta, M., Goodman, M., Cox, C. E., & Kip, K. E. (2009). Randomized controlled trial of mindfulness-based stress reduction (MBSR) for survivors of breast cancer. *Psycho-Oncology, 18*(12), 1261–1272. https://doi.org/10.1002/pon.1529

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198–1202.

Loh, W. (2011). Classification and regression trees. *Wires Data Mining and Knowledge Discovery, 1*(1), 14–23. https://doi.org/10.1002/widm.8

Luberto, C. M., Park, E. R., & Goodman, J. H. (2018). Postpartum outcomes and formal mindfulness practice in mindfulness-based cognitive therapy for perinatal women. *Mindfulness, 9*(3), 850–859. https://doi.org/10.1007/s12671-017-0825-8

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2019). *Consistent individualized feature attribution for tree ensembles*. Retrieved from https://arxiv.org/abs/1802.03888. Accessed 30 May 2021.

Lymeus, F., Lindberg, P., & Hartig, T. (2019). A natural meditation setting improves compliance with mindfulness training. *Journal of Environmental Psychology, 64*, 98–106. https://doi.org/10.1016/j.jenvp.2019.05.008

Manuvinakurike, R., Velicer, W. F., & Bickmore, T. W. (2014). Automated indexing of internet stories for health behavior change: Weight loss attitude pilot study. *Journal of Medical Internet Research, 16*(12), e285. https://doi.org/10.2196/jmir.3702

Maxhuni, A., Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O., & Morales, E. F. (2016). Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive and Mobile Computing, 31*, 50–66. https://doi.org/10.1016/j.pmcj.2016.01.008

Miller, Y. R., Medvedev, O. N., Hwang, Y.-S., & Singh, N. N. (2021). Applying generalizability theory to the perceived stress scale to evaluate stable and dynamic aspects of educators' stress. *International Journal of Stress Management, 28*(2), 147–153. https://doi.org/10.1037/str0000207

Molnar, C., Bischl, B., & Casalicchio, G. (2018). iml: An R package for interpretable machine learning. *Journal of Open Source Software, 3*(26), 786. https://doi.org/10.21105/joss.00786

Molnar, C. (2021). *Interpretable machine learning*. Retrieved from https://christophm.github.io/interpretable-ml-book/. Accessed 27 May 2021.

Morrison, L. G., Hargood, C., Pejovic, V., Geraghty, A. W. A., Lloyd, S., Goodman, N., Michaelides, D. T., Weston, A., Musolesi, M., Weal, M. J., & Yardley, L. (2017). The effect of timing and frequency of push notifications on usage of a smartphone-based stress management intervention: An exploratory trial. *PLoS ONE, 12*(1), e0169162. https://doi.org/10.1371/journal.pone.0169162

Mrazek, A. J., Mrazek, M. D., Cherolini, C. M., Cloughesy, J. N., Cynman, D. J., Gougis, L. J., Landry, A. P., Reese, J. V., & Schooler, J. W. (2019). The future of mindfulness training is digital, and the future is now. *Current Opinion in Psychology, 28*, 81–86. https://doi.org/10.1016/j.copsyc.2018.11.012

Parsons, C. E., Crane, C., Parsons, L. J., Fjorback, L. O., & Kuyken, W. (2017). Home practice in mindfulness-based cognitive therapy and mindfulness-based stress reduction: A systematic review and meta-analysis of participants' mindfulness practice and its association with outcomes. *Behaviour Research and Therapy, 95*, 29–41. https://doi.org/10.1016/j.brat.2017.05.004

Perich, T., Manicavasagar, V., Mitchell, P. B., & Ball, J. R. (2013). The association between meditation practice and treatment outcome in mindfulness-based cognitive therapy for bipolar disorder. *Behaviour Research and Therapy, 51*(7), 338–343. https://doi.org/10.1016/j.brat.2013.03.006

Quach, D., Gibler, R. C., & Jastrowski Mano, K. E. (2017). Does home practice compliance make a difference in the effectiveness of mindfulness interventions for adolescents? *Mindfulness, 8*(2), 495–504. https://doi.org/10.1007/s12671-016-0624-7

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier*. Retrieved from http://arxiv.org/abs/1602.04938. Accessed 17 Aug 2020.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., & Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing, 267*, 664–681. https://doi.org/10.1016/j.neucom.2017.06.053

Segal, Z., Dimidjian, S., Vanderkruik, R., & Levy, J. (2019). A maturing mindfulness-based cognitive therapy reflects on two critical issues. *Current Opinion in Psychology, 28*, 218–222. https://doi.org/10.1016/j.copsyc.2019.01.015

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology, 179*(6), 764–774. https://doi.org/10.1093/aje/kwt312

Shanmuganathan, S., & Samarasinghe, S. (Eds.). (2016). *Artificial neural network modelling*. Springer International Publishing. https://doi.org/10.1007/978-3-319-28495-8

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Tal, A., & Torous, J. (2017). The digital mental health revolution: Opportunities and risks. *Psychiatric Rehabilitation Journal, 40*(3), 263–265. https://doi.org/10.1037/prj0000285

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining, 10*(6), 363–377. https://doi.org/10.1002/sam.11348

Tierney, N., Di, C., McBain, M., & Fay, C. (2021). *naniar: Data structures, summaries, and visualisations for missing data* (R package version 0.6.1) [Computer software]. https://CRAN.R-project.org/package=naniar. Accessed 27 May 2021.

Titov, N., Hadjistavropoulos, H. D., Nielssen, O., Mohr, D. C., Andersson, G., & Dear, B. F. (2019). From research to practice: Ten lessons in delivering digital mental health services. *Journal of Clinical Medicine, 8*(8), 1239. https://doi.org/10.3390/jcm8081239

Toneatto, T., & Nguyen, L. (2007). Does mindfulness meditation improve anxiety and mood symptoms? A review of the controlled research. *The Canadian Journal of Psychiatry, 52*(4), 260–266. https://doi.org/10.1177/070674370705200409

Triantafyllidis, A. K., & Tsanas, A. (2019). Applications of machine learning in real-life digital health interventions: Review of the literature. *Journal of Medical Internet Research, 21*(4), e12286. https://doi.org/10.2196/12286

Von Neumann, J., Kent, R., Bellinson, H., & Hart, B. (1941). The mean square successive difference. *The Annals of Mathematical Statistics, 12*, 153–162.

Wahbeh, H., Svalina, M. N., & Oken, B. S. (2014). Group, one-on-one, or internet? Preferences for mindfulness meditation delivery format and their predictors. *Open Medicine Journal, 1*(1), 66–74. https://doi.org/10.2174/1874220301401010066

Yeo, C. J. J., Barbieri, A., Roman, G., Wiesman, J., & Powell, S. (2019). Using smartphone mindfulness apps to increase trainee resilience and reduce burnout. *Neurology, 92*(15 Supplement), P2.9–005

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 3*(1), 1–130. https://doi.org/10.2200/S00196ED1V01Y200906AIM006

Zhu, B., Hedman, A., Feng, S., Li, H., & Osika, W. (2017). Designing, prototyping and evaluating digital mindfulness applications: A case study of mindful breathing for stress reduction. *Journal of Medical Internet Research, 19*(6), e197. https://doi.org/10.2196/jmir.6955

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.