Do Exact Explanations Make a Difference? A Case Study Among Weight Management Experts

GLENN FERNANDES, Northwestern University
ARTHUR CHOI, Kennesaw State University
MAIA JACOBS, Northwestern University
ADNAN DARWICHE, University of California, Los Angeles
NABIL ALSHURAFA, Northwestern University

Increasingly, there is a growing body of literature supporting the use of machine learning models to advance domain experts' ability to predict and change the course of a person's health behavior. While machine-learned (ML) models can learn to predict human behavior, these learned ML models are often dubbed "black boxes," that are overly complex. This complexity impedes domain experts' model adoption, primarily due to the difficulty in interpreting, lack of face validity, or lack of concurrence with clinical knowledge. The result is that a domain expert might either overly rely on or mistrust the model's decisions. Traditional ML model evaluation metrics such as accuracy and sensitivity provide an overall metric of model performance, but fall short of providing insights into a model's prediction. Recent developments in Explainable ML has produced a generation of explanations (e.g., feature ranking, LIME, SHAPely values) that are based on approximations. However, advances using Prime Implicants enable us to develop exact explanations (i.e., local explanations that are faithful to the behavior of the original model). Although, these generated explanations might help users gain insight into the inner-workings of a model, it is unknown whether exact explanations improve trust and understanding of ML models for domain experts. In order to address this research question, we built an interactive tool called PRIMO (Prime Implicant Maintenance of Outcome) that allows weight management experts to generate exact explanations from an ML model generated to predict early, the success and failure of a weight-loss intervention. We present a study to evaluate the utility of explainability models in advancing trust among weight management experts, quantitatively, using a 'switch-to-agree' metric (based on the agreement of the domain-expert with the decisions of the machine learned model) and, qualitatively, through thematic analysis of responses to open-ended questions about trust and understanding. Our study promotes discussion on the utility of exact explanations in advancing model adoption, whether exact explanations advance model trust, whether the 'switch-to-agree' metric is ideal in capturing increased trust in model output, and best methods for communicating uncertainty of a models exact explanations. We also provide an end-to-end framework for the community to further test the effectiveness of computing exact explanations that are faithful to the behavior of the original model.

CCS Concepts: • **Computing methodologies → Machine learning approaches**; • **Human-centered computing → User studies**.

Additional Key Words and Phrases: machine learning, explainable, random forest, obesity

Authors' addresses: Glenn Fernandes, Northwestern University, glennfer@u.northwestern.edu; Arthur Choi, Kennesaw State University; Maia Jacobs, Northwestern University; Adnan Darwiche, University of California, Los Angeles; Nabil Alshurafa, Northwestern University, nabil@northwestern.edu.

Manuscript submitted to ACM

ACM Reference Format:
Glenn Fernandes, Arthur Choi, Maia Jacobs, Adnan Darwiche, and Nabil Alshurafa. 2018. Do Exact Explanations Make a Difference? A Case Study Among Weight Management Experts. *J. ACM* 37, 4, Article 111 (August 2018), 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Machine-learned (ML) models are used increasingly to understand health care decisions for the purpose of improving patient outcomes. ML models used to represent complex health behavior data are often black-boxes that are overly complex; for example, a Random Forest (RF) model—a classifier known for increasing predictive accuracy even without hyper-parameter tuning, comprising many decision trees where each tree uses a different number of features or variables to determine a classification—is difficult to understand intuitively. While RF outperforms several other ML methods in terms of prediction accuracy, it is notoriously hard to interpret [18]. There is an apparent trade-off between the performance of classifiers (accuracy) and their ability to explain the reasoning behind its results (explainability). Despite this, researchers have been quick to ride the predictability wave of RFs. Several publications have used RF algorithms for tasks such as risk prediction [29, 32], emotion detection [22], sign language recognition [25], predicting social interactions [8], gesture recognition [6], estimating energy prediction [20], and early detection of depression [3].

Machine learned models and explainability tools developed by computer scientists might help users gain insight into the inner workings of a model, but it is unclear whether health domain experts will agree with predictions provided by ML models. Visualization and interactive design research have refined methods to convey to domain experts the reasoning behind the decisions made by ML models. However, recent research has highlighted an over-reliance problem among clinicians in primary care settings, placing too much trust in automated systems can lead to misuse of a system where users agree with incorrect system suggestions. Jacobs et al. [12] recently reported that psychiatrists with higher familiarity with ML were less likely to use an ML recommendation of which antidepressant drug to use compared with clinicians with lower ML familiarity [12]. Although, the use of model explanations are expected to help mitigate the issue of over-reliance, since health behavior is often difficult to predict, it is unknown whether there would be lack of adoption of ML or over-reliance among health behavior experts.

Traditional ML evaluation metrics such as accuracy and sensitivity provide an overall metric of the model's performance but fall short of providing insights into the model's prediction. Recent developments in the field of explainable artificial intelligence (AI) have seen new explainability metrics or explanations that provide further understanding about the reasoning behind the decisions made by an ML model. One such example of an explainability metric, or explanation, is feature ranking to understand the degree to which features influence a model. Model-agnostic explanations such as SHAP [16], LIME [23], and ANCHORS [24] are currently the standard approaches to explainability. LIME and ANCHORS rely on local explanations that are aimed at explaining the model's reasoning for a given instance (eg, to answer the question, Why was this instance classified as "X?"); however, these algorithms are based on approximations by defining the contribution of a feature to the difference between the actual and mean prediction (ie, they are not accurate explanations). Further, they are not sparse explanations, preferred by humans, and can not be used to make statements about changes in prediction as input changes. SHAP has a solid foundation in game theory and provides both consistent global and local explanations; however it is either slow (to generate explanations) and ignores feature dependence or it produces unintuitive feature attributions [16].

Recent advances in the ability to compile classifiers into decision graphs have enabled the potential to generate what we call *exact explanations*, which are explanations that are faithful to the behavior of the original model. Shih et al. and others [5, 27, 28] designed methods that compile ML models, such as an RF classifier into ordered decision graphs, a tractable representation, also called ordered binary decision diagrams (OBDDs). OBDD are tractable representations of a ML model that enable researchers to perform queries, such as computing prime implicants (minimal set of feature values that are responsible for a decision), on the learned model to facilitate greater understanding. However, it is unclear whether these exact explanations will advance model adoption among end users, and whether they will have an effect on model trust.

Designing an interface where end users, particularly domain experts, can interact with the ML models and its explanations is essential to building model trust. As shown by Wang et al. [30], co-design with end users provides the ability to understand, first hand, if a theoretical approach to generating explanations of a certain type, including the design of the interactive tool to visualize them, will translate well among domain-experts. We use this tractable representation alongside visualization methods used in prior research to build PRIMO (Prime Implicant Maintenance of Outcome), a first attempt at building a tool to generate, visualize, and interact with prime implicant–based explanations.

In this paper we discuss how to build a weight-loss prediction model using the Random Forest classifier. We then discuss PRIMO, a software tool and pipeline for converting an RF model into a tractable representation, providing exact explanations. We then discuss a study design intended to assess ML model adoption among weight management experts and to assess the impact of using PRIMO on ML model trust. In doing so we aim to promote discussion on the utility of exact explanations in advancing model adoption, whether exact explanations advance model trust, optimal metrics for capturing increased trust in model output, and best methods for communicating uncertainty of a models exact explanations.

2 BACKGROUND AND RELATED WORKS

Explainable AI approaches can be grouped based on the type of explanation and the underlying techniques used to generate those explanations.

2.1 Local and Global Explanations

Local explanations are those aimed at explaining the model's reasoning for a given instance [17]. Rebeiro et al. [23] designed model-agnostic local explanation techniques to explain the decisions of any classifier. Their approach, called LIME (Local Interpretable Model-Agnostic Explanations), is based on learning the local behavior of a complex classifier (e.g., a random forest) around a given instance, using a simpler but interpretable classifier (e.g, using a decision tree). The accuracy of this approximation depends primarily on the ability of the simpler classifier to (locally) approximate the original model's decision boundaries. To generate local explanations that are more precise, Reberio et al. take another approach by solving a multi-armed bandit problem to generate interpretable if-then rules. This technique overcomes the issue of local complexity but at the cost of over-fitting explanations locally. Although local explanations are specific to understanding a specific instance, when combined together with explanations of multiple instances, it could provide an understanding of the model's overall behavior, similar to a global explanation. Global explanations are explanations that describe the overall working of the ML model [17], such as feature importance. Lundberg et al.'s SHAP [16] can be used to create global explanations by aggregating the Shapley values to create feature importance, summary, and dependence plots. Shapley values are feature attributions that act as driving forces either contributing to the prediction or not. This implies that, unlike LIME, SHAP does not train an interpretable model that can make predictions.

2.2 Decision rule based

If-then decision rule-based global explanations such as Bayesian decision lists (BDL) [15] strike a balance between accuracy and interpretability. Although BDL explanations seem similar to those by Ribeiro et al. [24], BDL is a global explanation whereas anchors is local (explaining a specific instance), and these explanations also differ in the techniques used to generate them. Lakkaraju et al.'s [13] decision sets are an unordered set of decision lists aimed at increasing user interpretability over decision lists, and their user study shows that participants were able to describe decision sets much faster than decision lists. To further increase user interpretability Lakkaraju et al. [14] incorporated user inputs and generated a subspace of decision set explanations tailored to the users' interests. These decision rules-based explanations are interpretable classifiers in it of themselves and are not model agnostic.

2.3 Design Principles

Recently, Melis et al. proposed a list of design principles to keep in mind while designing explanations, based on philosophy, cognitive science and social science. These principles state that explanations should be designed such that they are contrastive, exhaustive, modular and compositional, easlily understandable quantities, and parsimonius. To generate exact explanations that aim at following the above principles, we planned to take a unique approach of generating explanations that can integrate local and global aspects while ensuring user-specific interest. Our approach looks at converting a Random Forest model into logic.

2.4 Trust and Reliance

An intelligible system is one that is understandable by the user and predictable through the use of explanations. One practical approach to developing intelligible systems is through explainable AI by creating explanations that adhere to user-desired properties of understandability and predictability, as well as user-desired outcomes of trustworthiness, reliability, and safety. Trust has been operationalized through constructs of confidence, reliability, predictability, and efficiency [19],[4]. Cahour et al. [4] used these constructs in the form of a Likert scale to evaluate trust in a cruise control system. We adapted these questions in our quantitative and qualitative interviews to evaluate domain experts' trust and reliance in our system. Bussone et al. [2] measured primary care practitioners' trust and reliance in a clinical decision support system by measuring how many participants agreed or disagreed with the explanation and by also measuring the difference in trust ratings (7-point Likert scale) before and after showing confidence explanations and "why" explanations. The authors concluded that showing a "why" explanation leads to over-reliance on system suggestions—even if the suggestions are incorrect. Ribeiro et al. [23] measured users' trust in their local, interpretable model-agnostic explanations generated on a biased data set and interviewed users after they made their own predictions against the classifiers in the study. The idea of measuring what proportion of users that agreed or switched their prediction based on observing a model's prediction and explanation was implemented by Yin et al. [31] to analyze how users' perception of a model's accuracy impacts trust.

3 METHODS

Our proposed framework consists of three steps: (1) Selecting an optimal early prediction time point, (2) generating a tractable decision diagram representation of the RF model using the RF classifier, and (3) Building a software tool to enable visualization and interaction with the explanations. We describe each of these steps below, followed by details of our one-on-one interviews with domain experts using our software tool, the statistical analysis techniques to assess agreement of health domain experts with our ML model output, and the questions used to assess trust and understanding of ML models.

3.1 Explainability Definition

Explainable AI involves the communication of ML model results and operations for different audiences and purposes, and our main goal is to study its effectiveness in communicating these explanations to the health domain expert community. To do this we define explainability as the ability for domain experts to trust and understand the explanations presented when related to the problem at hand.

3.2 Weight Loss Study

The problem we have identified to explore is early-prediction of weight loss. Obesity-attributable medical expenditures remain high, but interventions effective and economical have not been adequately identified. Predicting the likelihood of success of weight loss in interventions using machine learning (ML) models may enhance intervention effectiveness by enabling timely and dynamic modification of intervention components. However, lack of understanding and trust of these methods impacts adoption among weight management experts. Developments in explainable ML techniques enable the generation of explanations to interpret decisions of ML models, yet it is unknown how to enhance model understanding and trust among weight management experts.

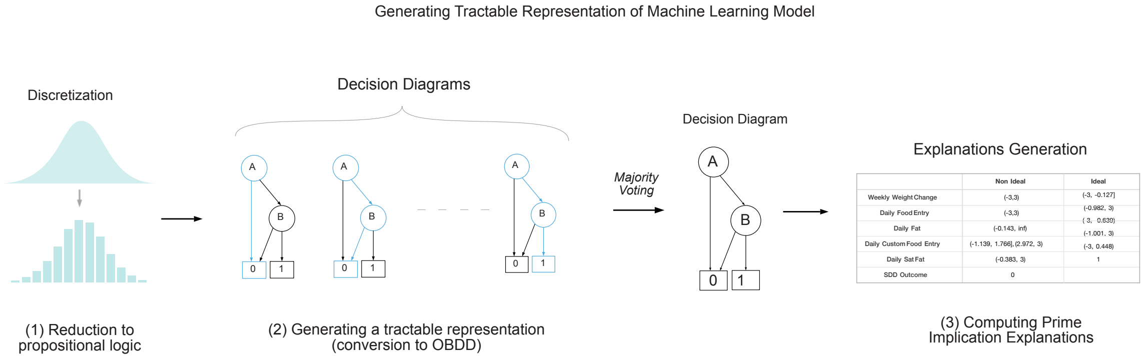
We used the Opt-IN dataset to train Random Forest Models for early-prediction of weight loss. The Opt-IN weight-loss study was a 6-month, theory-guided, technology-supported weight loss intervention aimed at exploring factors that contribute to improvement in achieving meaningful weight loss. Details of the study have been previously described [22,23] .The study enrolled adults with body mass index between 25.1 and 39.9 kg/m2. Eligible participants had stable weight (no loss or gain>25 lbs for the past 6 months), were not enrolled in any formal weight loss program or taking weight-reducing medications and were interested in losing weight. Participants obtained their personal physician's approval to participate, and the physician agreed to receive study reports. All study procedures were approved by the Institutional Review Board, and all participants provided written informed consent prior to enrolment. We used demographics information and data collected form participant smartphones to determine what factors early in the intervention predicted weight loss at 6 months. We define clinically meaningful weight loss to be at least 7% weight reduction from baseline.

3.3 Framework

3.3.1 Building RF model at an optimal early prediction time point. In order to identify the critical early time-point for building a machine-learned model, we built several learned models at different time points to select the model-time-point pair with the highest predictability. We combined both evidence-based and data-driven practices to guide the process. Evidence-based practice guided our initial selection of features and the subset of time points for early prediction to select from. Data-driven approaches guided the dimensionality reduction through feature selection and development of machine-learned model at each time-point. We identified a minimal number of highly predictive features combining self-reported dietary (initial weight loss, fat intake, saturated fat intake) and engagement variables (entry of food items, entry of custom food items) to build the RF models. We built multiple RF classifiers (using the OPT-IN dataset, 419 participants after data cleaning and filtering) to predict a binary outcome of weight loss at the end of 26 weeks. We observed a local optimum in predictability of models at the end of week 2 and at the end of week 3. We decided to select an earlier timepoint and therefore selected the end of week 2 as the optimum timepoint. We selected the best RF Model from this timepoint with 13 trees, depth 3 and accuracy 81% to convert to logic.

3.3.2 Generating a tractable representation to facilitate computation of prime implicant explanations The random forest model underwent three steps. [Figure 1]

(1) Reduction to propositional logic: A random forest classifier generally takes continuous variables as input, but each continuous feature becomes a proposition when it appears in any given decision tree. Hence, we can view the input/output behavior of any random forest classifier as a propositional function. A random forest classifier consists of an ensemble of decision tress, and it classifies an instance by evaluating each decision tree. Note that a random forest classifier (and each decision tree in its ensemble) evaluates instances using variable tests of the form xi >= t or xi < t, which are both propositions. When we take all of the variable tests for a particular feature Xi, they induce a partitioning of the space of Xi into mutually exclusive and exhaustive intervals. In our reduction, we represent each interval as a binary variable in our propositional formula, which can be viewed as a discretization of the original continuous variable.


Generating Tractable Representation of Machine Learning Model
Discretization
Decision Diagrams
A
B
0
1
Majority Voting
Decision Diagram
A
B
0
1
Explanations Generation
(1) Reduction to propositional logic
(2) Generating a tractable representation (conversion to OBDD)
(3) Computing Prime Implication Explanations

	Non Ideal	Ideal
Weekly Weight Change	(-3,3)	(-3, -0.127]
Daily Food Entry	(-3,3)	(-0.982, 3)
Daily Fat	(-0.143, inf)	(-3, 0.639) (-1.001, 3)
Daily Custom Food Entry	(-1.139, 1.766], (2.972, 3)	(-3, 0.448)
Daily Sat Fat	(-0.383, 3)	1
SDD Outcome	0	

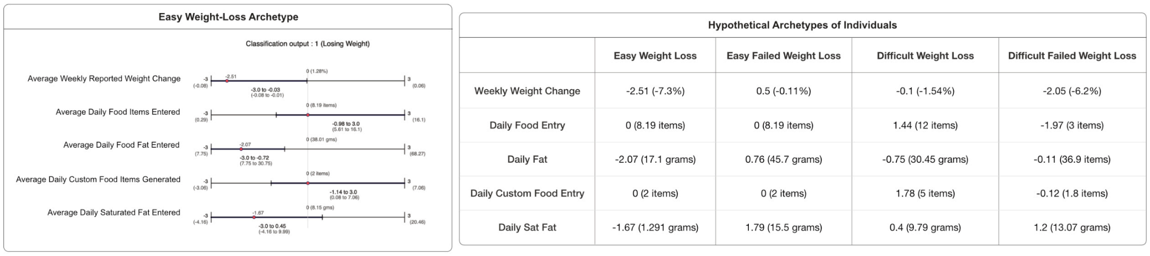
Fig. 1. Generating Tractable Representation of Machine Learning model to compute prime implicant explanations

(2) Generating Tractable Representation (Conversion to OBDD): Simply obtaining the propositional formula underlying a random forest is not useful, since reasoning with the formula will not be tractable. For example, testing whether there exists a satisfying assignment (i.e., testing whether there exists some feature vector that obtains a given label) is an NP-hard problem. Hence, we appeal to the field of knowledge compilation to obtain a tractable representation of the formula [24]. Knowledge compilation is a sub-field of AI that studies in part, tractable Boolean circuits, and the trade-offs between succinctness and tractability. That is, by enforcing different properties on the structure of a Boolean circuit, one can obtain greater tractability (the ability to perform certain queries and transformations ifn polytime) at the possible expense of succinctness (the size of the resulting circuit). We follow Choi et al [20] in order to compile the propositional formula of a random forest into an ordered decision graph, or equivalently, an OBDD. Once we compile a propositional formula into an OBDD, many queries of interest become tractable, typically requiring time that is only linear in the size of the resulting OBDD. We describe some of these queries next.

(3) Computing prime-implicant explanations: To generate explanations, we compiled the discrete random forest classifier into an OBDD, which is a tractable representation of a function and can be used to efficiently answer queries and facilitate efficient explanations of classifiers. Once we have an OBDD representation of our random forest classifier, we reason about and generate explanations for the behavior of the classifier [25,26]. One type of explanation is called

a sufficient explanation [27] which corresponds to computing the prime implicants of an random forest classifier's propositional function. A prime implicant of a random forest's propositional formula can be thought of as a minimal assignment of features to values that will fix the output of the random forest classifier.

For each instance, that is, each set of input to the classifier, we generated a shortest prime implicant that is compatible with it (i.e., the shortest sub-instance that is also a prime implicant). As explained by Shih et al. [26]one way of verifying the behavior of a classifier, is to verify whether the classifier is compatible with the expectations of a domain expert. A domain expert may define expectations of input-output pairs for a classifier to be reliable. For example, a domain expert may say that anyone who has lost at least 1% of their weight and is maintaining low intakes of saturated fat early in treatment, is on the trajectory toward clinically meaningful distal weight loss. To facilitate this understanding of expected behavior, we provide a visual of intervals/ranges for each variable, such that if the value of each variable falls within these ranges, the classifier output will remain the same. Domain experts need the ability to visualize and understand explanations quickly and intuitively. We designed a visual for the explanation that highlights the ranges for each of the variables on a number line. The visual also shows the z-score values in addition to the actual scale for each of the features. Figure 2 shows a sample instance of an easy weight-loss archetype and it's corresponding visual, which we call Prime Implicant Maintenance of Outcome (PRIMO).



Hypothetical Archetypes of Individuals

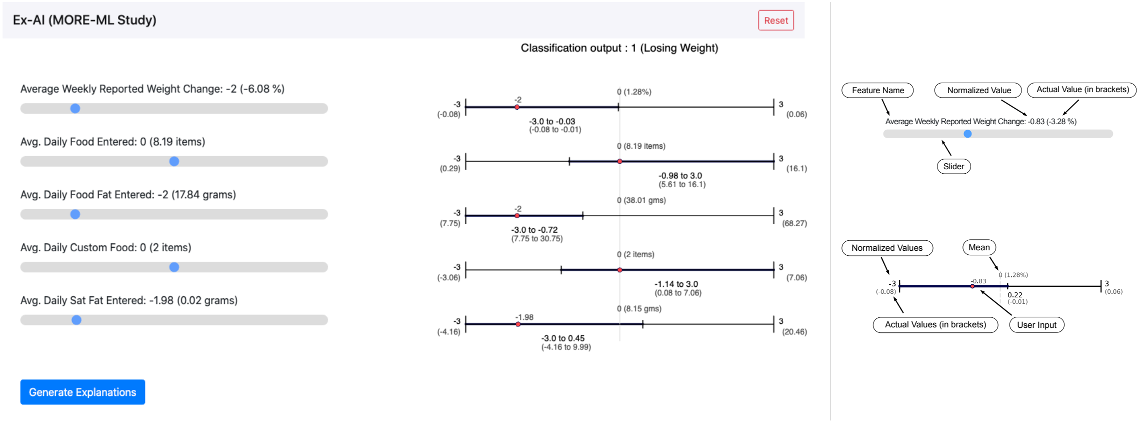
	Easy Weight Loss	Easy Failed Weight Loss	Difficult Weight Loss	Difficult Failed Weight Loss
Weekly Weight Change	-2.51 (-7.3%)	0.5 (-0.11%)	-0.1 (-1.54%)	-2.05 (-6.2%)
Daily Food Entry	0 (8.19 items)	0 (8.19 items)	1.44 (12 items)	-1.97 (3 items)
Daily Fat	-2.07 (17.1 grams)	0.76 (45.7 grams)	-0.75 (30.45 grams)	-0.11 (36.9 items)
Daily Custom Food Entry	0 (2 items)	0 (2 items)	1.78 (5 items)	-0.12 (1.8 items)
Daily Sat Fat	-1.67 (1.291 grams)	1.79 (15.5 grams)	0.4 (9.79 grams)	1.2 (13.07 grams)

Fig. 2. In the above PRIMO visual, for easy weight-loss archetype, the red point indicates the patient profile. The easy weight-loss archetype tends to be easy because the red points fall well within the edges of the PRIMO generated prime implicant ranges. The table on the right, shows the different hypothetical archetypes of individuals.

3.4 Designing a tool for weight-management experts

An Interactive software tool was designed to enable researchers to query the ML model and generate explanations for hypothetical archetypes of individuals. The user can create custom instances by assigning values to each feature and then select the "Generate Explanation" button [Figure 3]. The software queries the interpretable model to provide a prediction and PRIMO's visual. PRIMO was designed to be interactive and intuitive by enabling users to enter values for features of the model by adjusting sliders that represent standardized values for each feature. The sliders show both the z-score and corresponding actual value for each of the features to create an instance.

The explanation provided by the model is displayed in the form of highlighted ranges on the interface. If one were to tweak the input values of an instance within the respective ranges specified by the explanation, the model's prediction and explanation is guaranteed to not change. However, on tweaking the input value outside the displayed ranges, the prediction and the explanation may or may not change. This provides users with the ability to understand the limits of the model that may or may not produce a different output, providing better insight into the models functionality. In some cases, there is a possibility that, on selecting certain inputs, there is no explanation range for one or more features.


Ex-AI (MORE-ML Study)

Classification output : 1 (Losing Weight)

Average Weekly Reported Weight Change: -2 (-6.08 %)

Avg. Daily Food Entered: 0 (8.19 items)

Avg. Daily Food Fat Entered: -2 (17.84 grams)

Avg. Daily Custom Food: 0 (2 items)

Avg. Daily Sat Fat Entered: -1.98 (0.02 grams)

Fig. 3. Software Interface

This indicates that given the values for all the features with ranges defined, the value of the feature without a range does not affect the output of the predictor.

3.5 Study Design for Evaluating Trust and Reliance with Domain-Experts

The primary goal of the study was to evaluate among end users', the weight management experts, if the explanations are understandable and if they would trust to use this tool in a real-world scenario. The participants in the study include weight-management experts with backgrounds in one or more of the following areas: psychology, nutrition, epidemiology and clinical experience, statistics, and data science.

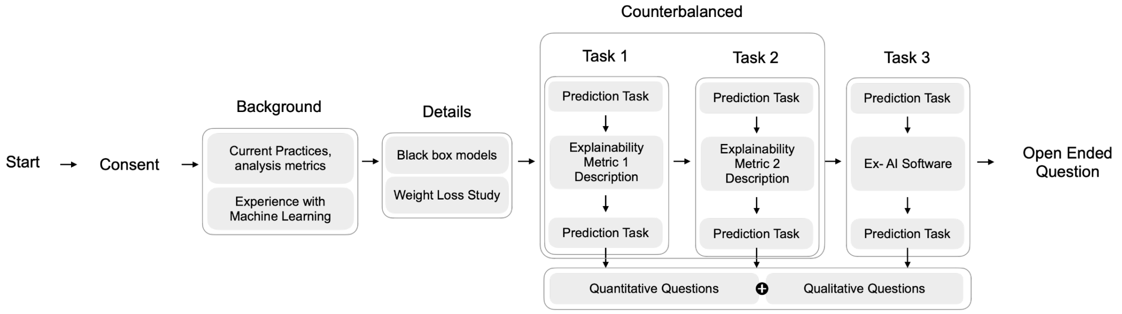


Fig. 4. Study Design

The participants in the study [Figure 4] were asked about their background and their experience with analysis metrics and ML. They were presented with details of the weight loss intervention for which the random forest model was based. They were then required to provide feedback on 2 other explanation types in addition to PRIMO. These 2 explanation types were used to help participants get acquainted with the idea of explainability and using explainability tools to understand predictions form a ML classifier. Our explainability framework enabled the generation of these other 2 explanation types as well, which closely resembles feature ranking based and conditional probability based explanation types. For each metric, we created four hypothetical archetypes of individuals [Figure 2], two easy and two difficult cases for prediction, including: (1) a participant expected to lose sufficient weight at the end of the study (easy weight loss); (2) a participant expected to not lose weight (easy failed weight-loss); (3) a participant that was borderline but expected to lose sufficient weight (difficult weight loss); and (4) a participant that was borderline and not expected to lose weight (difficult failed weight loss).

3.5.1 Defining Trust and Measuring Trust Quantitatively. We define trust to be the domain experts' tendency to agree with the model's explanations and thus with the prediction.

We asked participants to predict the outcome of each hypothetical archetype before and after viewing an explanation.

The quantitative questions were evaluated on a Likert response scale ranging from "not at all" (0), "a little" (1), "somewhat" (2), "quite a lot" (3), and "extremely" (4). The five quantitative questions were presented as follows: (adapted from Cahour et al. [30]) (1) How much do you trust the model? (2) How reliable do you think the explainability metric is? (3) According to you, how predictable are the outputs of this model? (4) According to you, how efficient is the explainability metric in describing why the model generated the outputs/predictions? and (5) How confident are you in your answers?

We then calculated and compared the mean (SE) of the Likert response scale values answered by the participants for the five quantitative questions shown after each task.

We defined trust to be the domain experts' tendency to agree with the model's explanations and thus with the prediction. To measure this quantitatively, based on the definition of agreement fraction and switch fraction by Yin et al. [31], we define three metrics: switch-to-agree, switch-to-disagree and switch-ratio. Participants were required to make predictions for the four hypothetical archetypes before and after going over each explainability metric. Across metrics, a different set of values for the four archetypes was shown; however, for each individual metric the same set of values for the four archetypes was shown to the participant before and after going over the explainability metric. A switch occurs when a participant changes their prediction for an archetype after seeing the explainability metric. Since there are 4 archetypes and 14 participants, the maximum number of switches possible would be 14 * 4 = 56. We defined the following evaluation metrics based on participants' agreement or disagreement with the ML model's decision:

1. Switch-to-agree: number of times participants switched their response towards a prediction in agreement with the ML model's decision.
2. Switch-to-disagree: number of times participants switched their response towards a prediction that differed from the machine learning model's decision.
3. *Switch Ratio:* The proportion of switch-to-agree to switch-to-disagree.

3.5.2 Analyzing Trust Qualitatively. Qualitative questions were designed to engage the user in an open-ended response to capture how and why the metric facilitates prediction of outcomes. The three open-ended questions were presented as follows: (1) Would you trust this model guided by this evaluation metric to work well in the real world? (adapted from Ribeiro et al. [16]) (2) How do you think the model is able to distinguish between the classes? (adapted from Ribeiro et al. [16]) (3) Do you have a better understanding of the model? If yes, why? If no, how would you improve the explainability metric? (Open Ended). At the end of the study the participants were asked to rank the three explanation types and to describe their suggestions regarding improvements and which types enhance their understanding of ML models. We also report on themes regarding participant preference for a certain type of metric over another by qualitatively analyzing their responses.

4 DISCUSSION

4.1 Utility of exact explanations in advancing model adoption

Given we are the first to study model adoption among weight management experts, we hope to promote discussion in the rigor of our study design. In doing so we hope to understand the limitations of our study design, and discuss potential ways of strengthening the study design to gain further insight into ML model adoption among weight management experts.

4.2 Measuring ML model trust and reliance

A switch-to-agree scenario indicates that the explanation had an effect on the person's decision making, and increased their tendency to agree with the explanation and decision of the model after interacting with PRIMO. For a given explanation type if the number of switch-to-agree cases were high among participants, it indicates that the explanation type instills a change in behavior and decision making on the participant's end, there-by increasing their trust in the original model. One might argue that this metric does not capture moments where participants agreed with the model prior to viewing the explanation. In this scenario it is difficult to determine the effect of the explanation, reducing confidence in the switch-to-agree metric capturing reliance on the explanation metric, and thus, further strengthens the need to ask participants more specific qualitative questions to gauge reliance. The same argument can be made for the switch-to-disagree metric.

To further improve the quantification of trust, we measure the proportion of switch-to-agree to switch-to-disagree cases for a given explanation type, among participants in the study. A high switch-ratio indicates a high switch-to-agree and low switch-to-disagree, which implies the explanation increased potential for agreement. However, it is unclear whether this metric is sufficient to measure trust and reliance. To understand participants' trust and reliance we directly ask the participant how much they trust the model, and if the explanation was reliable for them to make a decision. These open-ended type questions may allow us to further investigate how PRIMO may be optimized to increase trust and reliance.

4.3 Communicating uncertainty of a model's exact explanation

Initial interviews with weight management experts are showing the need to visualize uncertainty in the explanations. Our use of exact explanations minimizes discrepancies between the behavior of the explainable model and original model. This implies measuring uncertainty in the decisions of the original model is equivalent to measuring uncertainty in the decisions of the explainable model and an effective approach to communicate uncertainty is by visualizing it. One way to visualize uncertainty is to visualize the dataset the model was trained on, and highlight the differences between the ground truth and performance of the explanation ranges on the original dataset. Several studies have shown icon-arrays to be effective in conveying information related to risk statistics [9–11, 33]. Icon-arrays are also used to communicate risks and uncertainty in several domains [1, 7, 26], because it is not only effective for communicating information related to comparison of ratios, but also highly effective in conveying uncertainty information related to frequency framing [21]. Therefore, we plan to use icon-arrays in the future to visualize patient-outcomes in the weight-loss study, to enhance PRIMO's visual. In addition to viewing the data, providing clinicians the ability to filter data and select ranges for each of the features will enable them to create hypothetical profiles to identify patterns or outliers among the patients. However, selecting ranges for the individual features without knowing the distribution for each of the features can possibly lead to further uncertainty. This can be accounted for by incorporating density plots under each of the sliders, to visualize the distribution of each of the features. We hope to promote further discussion in methods of communicating uncertainty while using exact explanations.

5 NEXT STEPS

Through this research we have developed new methods for generating exact explanations that are faithful to the behavior of the original model. We are currently working on refining and testing the framework. We aim to share this approach with workshop participants and discuss the utility of exact explanations for health domain experts. Such discussions will be particularly important for considering the generalizability of this approach and the role exact explanations may play in establishing model adoption and trust. We look forward to feedback from the workshop to incorporate into our study design and analysis to improve our understanding of the utility of exact explanations and model adoption among health domain experts.

REFERENCES

[1] Paul L Aronson, Mary C Politi, Paula Schaeffer, Eduardo Fleischer, Eugene D Shapiro, Linda M Niccolai, Elizabeth R Alpern, Steven L Bernstein, and Liana Fraenkel. 2021. Development of an app to facilitate communication and shared decision-making with parents of febrile infants 60 days old. *Academic Emergency Medicine* 28, 1 (2021), 46–59.

[2] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[3] Fidel Cacheda, Diego Fernandez, Francisco J Novoa, Victor Carneiro, et al. 2019. Early detection of depression: social network analysis and random forest techniques. *Journal of medical Internet research* 21, 6 (2019), e12554.

[4] Béatrice Cahour and Jean-François Forzy. 2009. Does projection into use improve trust and exploration? An example with a cruise control system. *Safety science* 47, 9 (2009), 1260–1270.

[5] Arthur Choi, Andy Shih, Anchal Goyanka, and Adnan Darwiche. 2020. On Symbolically Encoding the Behavior of Random Forests. In *3rd Workshop on Formal Methods for ML-Enabled Autonomous Systems (FoMLAS)*.

[6] Klen Čopič Pucihar, Christian Sandor, Matjaž Kljun, Wolfgang Huerst, Alexander Plopski, Takafumi Taketomi, Hirokazu Kato, and Luis A Leiva. 2019. The missing interface: micro-gestures on augmented objects. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[7] Kyle W Davis, Debra L Roter, Tara Schmidlen, Laura B Scheinfeldt, and William MP Klein. 2021. Testing a best practices risk result format to communicate genetic risks. *Patient Education and Counseling* 104, 5 (2021), 936–943.

[8] Julian Frommel, Valentin Sagl, Ansgar E Depping, Colby Johanson, Matthew K Miller, and Regan L Mandryk. 2020. Recognizing affiliation: Using behavioural traces to predict the quality of social interactions in online games. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16.

[9] Mirta Galesic, Rocio Garcia-Retamero, and Gerd Gigerenzer. 2009. Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychology* 28, 2 (2009), 210.

[10] Rocio Garcia-Retamero and Mirta Galesic. 2010. Who proficts from visual aids: Overcoming challenges in people's understanding of risks. *Social science & medicine* 70, 7 (2010), 1019–1025.

[11] Rocio Garcia-Retamero, Mirta Galesic, and Gerd Gigerenzer. 2010. Do icon arrays help reduce denominator neglect? *Medical Decision Making* 30, 6 (2010), 672–684.

[12] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry* 11, 1 (2021), 1–9.

[13] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[14] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 131–138.

[15] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics* 9, 3 (2015), 1350–1371.

[16] Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).

[17] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2018. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *arXiv preprint arXiv:1811.11839* (2018).

[18] Toshiki Mori and Naoshi Uchihira. 2019. Balancing the trade-off between accuracy and interpretability in software defect prediction. *Empirical Software Engineering* 24, 2 (2019), 779–825.

[19] Bonnie M Muir. 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 11 (1994), 1905–1922.

[20] Ruairi O'Driscoll, Jake Turicchi, Mark Hopkins, Cristiana Duarte, Graham W Horgan, Graham Finlayson, R James Stubbs, et al. 2021. Comparison of the Validity and Generalizability of Machine Learning Algorithms for the Prediction of Energy Expenditure: Validation Study. *JMIR mHealth and uHealth* 9, 8 (2021), e23938.

[21] Lace Padilla, Matthew Kay, and Jessica Hullman. 2020. Uncertainty visualization. (2020).

[22] Prateek Panwar and Christopher M Collins. 2018. Detecting negative emotion for mixed initiative visual analytics. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[25] Jacob Schioppo, Zachary Meyer, Diego Fabiano, and Shaun Canavan. 2019. Sign Language Recognition: Learning American Sign Language in a Virtual Environment. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[26] Claudia R Schneider, Alexandra LJ Freeman, David Spiegelhalter, and Sander van der Linden. 2021. The effects of quality of evidence communication on perception of public health information about COVID-19: two randomised controlled trials. *medRxiv* (2021).

[27] Weijia Shi, Andy Shih, Adnan Darwiche, and Arthur Choi. 2020. On Tractable Representations of Binary Neural Networks. In *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR)*.

[28] Andy Shih, Arthur Choi, and Adnan Darwiche. 2019. Compiling Bayesian network classifiers into decision graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7966–7974.

[29] Svitlana Surodina, Ching Lam, Svetislav Grbich, Madison Milne-Ives, Michelle van Velthoven, Edward Meinert, et al. 2021. Machine Learning for Risk Group Identification and User Data Collection in a Herpes Simplex Virus Patient Registry: Algorithm Development and Validation Study. *JMIRx Med* 2, 2 (2021), e25560.

[30] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[31] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[32] Junsang Yoo, Si-Ho Kim, Sujeong Hur, Juhyung Ha, Kyungmin Huh, Won Chul Cha, et al. 2022. Correction: Candidemia Risk Prediction (CanDETEC) Model for Patients With Malignancy: Model Development and Validation in a Single-Center Retrospective Study. *JMIR Medical Informatics* 10, 1 (2022), e36385.

[33] Brian J Zikmund-Fisher, Holly O Witteman, Mark Dickson, Andrea Fuhrel-Forbis, Valerie C Kahn, Nicole L Exe, Melissa Valerio, Lisa G Holtzman, Laura D Scherer, and Angela Fagerlin. 2014. Blocks, ovals, or people? Icon type affects risk perceptions and recall of pictographs. *Medical decision making* 34, 4 (2014), 443–453.