Physiological Indicators for User Trust in Machine Learning with Influence Enhanced Fact-Checking

Jianlong Zhou[1]([✉]), Huaiwen Hu[2], Zhidong Li[1], Kun Yu[1], and Fang Chen[1]

[1] University of Technology Sydney, Sydney, Australia
{jianlong.zhou,zhidong.li,kun.yu,fang.chen}@uts.edu.au
[2] Data61, CSIRO, Sydney, Australia
vivianhhw@gmail.com

Abstract. Trustworthy Machine Learning (ML) is one of significant challenges of "black-box" ML for its wide impact on practical applications. This paper investigates the effects of presentation of influence of training data points on machine learning predictions to boost user trust. A framework of fact-checking for boosting user trust is proposed in a predictive decision making scenario to allow users to interactively check the training data points with different influences on the prediction by using parallel coordinates based visualization. This work also investigates the feasibility of physiological signals such as Galvanic Skin Response (GSR) and Blood Volume Pulse (BVP) as indicators for user trust in predictive decision making. A user study found that the presentation of influences of training data points significantly increases the user trust in predictions, but only for training data points with higher influence values under the high model performance condition, where users can justify their actions with more similar facts to the testing data point. The physiological signal analysis showed that GSR and BVP features correlate to user trust under different influence and model performance conditions. These findings suggest that physiological indicators can be integrated into the user interface of AI applications to automatically communicate user trust variations in predictive decision making.

Keywords: Influence · Machine Learning · Trust · Physiological features

1 Introduction

We have witnessed a rapid increase in the availability of data sets in various fields, for example in infrastructure, transport, energy, health, education, telecommunications, and finance. Together with the dramatic advances in Machine Learning (ML), getting insights from these "big data" and data analytics-driven solutions are increasingly in demand for different purposes. While we continuously

find ourselves coming across ML-based Artificial Intelligence (AI) systems that seem to work or have worked surprisingly well in practical scenarios (e.g. the self-driving cars, and the conversational agents for self-services), ML technologies still face prolonged challenges with low user acceptance of delivered solutions as well as seeing system misuse, disuse, or even failure. These fundamental challenges can be attributed to the nature of the "black-box" of ML methods for domain experts when offering ML-based solutions [36]. For example, for many non-ML users, they simply provide source data to an AI system, and after selecting some menu options, the system displays colorful viewgraphs and/or recommendations as output [37]. It is neither clear nor well understood *why ML algorithms made this prediction*, or *how trustworthy this output or decision based on the prediction was*. These questions demonstrate that both the explanation of and trust in ML play significant roles in affecting the user acceptance of ML in practical applications. The explanation is closely related to the concept of interpretability, which is referred to as the ability of an agent to explain or to present its decision to a human user, in understandable terms [5,29]. Trust is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [22].

As a result, recent research suggests model *interpretability/explanation* as a remedy for the "black-box" ML methods [24,26,36]. While there is much work in progress towards improving ML interpretability [15,18,30], the ideal state of having explainable, evidence driven ML-based decisions still remains a challenge [26]. To date, most of the work on ML interpretability has focused explicitly on ML model explanation itself, developing various explanation approaches to show why a prediction is made. However, the ML model explanation is just one component of the ML pipeline. Furthermore, what and how explanation information are presented to end users for the deployment to boost user trust plays significant roles in an ML-based intelligent (AI) system. Taking the influence of training data points on predictions [18] in supervised learning as an example, the explanation with influence allows to capture the weight/contribution of each training data point on the prediction of a testing data point. However, these explanations are highly biased towards ML experts' views, and are largely dependent on abstract statistical algorithms, which introduce further complexities to domain users. While domain users are more interested in what influence information affect and how these influence information are presented to them to boost their trust in predictions or decisions based on predictions.

Therefore, besides explanation, Mannarswamy et al. [26] proposed that the ability to provide justifiable and reliable evidences for ML-based decisions would increase the trust of users. Yin et al. [32] found that the stated model accuracy had a significant effect on the extent to which people trust the model, suggesting the importance of communication of ML model performance for user trust. Recently, *fact-checking*, which provides "evaluation of verifiable claims made in public statements through investigation of primary and secondary sources" [19], is increasingly used to check and debunk online information because of credibility challenges of the internet content [7]. Furthermore, previous research found the physiological correlations with decision making [39], it is possible that user

trust in predictive decision making can be evaluated by monitoring specific physiological signals for intelligent user interface of AI applications.

Motivated by these investigations, this paper introduces *fact-checking* into ML explanation by referring training data points as facts to users to boost user trust. These training data points are selected based on their influence level of predictions. We aim to investigate what influence of training data points and how they affect user trust in order to enhance ML explanation and boost user trust. We tackle this question by allowing users check the training data points that have the higher influence on the prediction and the training data points that have the lower influence on the prediction. The model performance is also introduced into the pipeline to find how both the influence and model performance affect user trust. Physiological signals are also collected and analysed to find their correlations to trust under different influence and model performance conditions.

2   Related Work

2.1   Explanation for Machine Learning

In the early years, visualization is primarily used to explain simple ML algorithms. For example, different visualization methods are used to examine specific values and show probabilities of selected objects visually for Naïve-Bayes [4], decision trees [2], or SVMs [9]. Advanced visualization techniques are then used as an interaction interface for users in data analysis. Guo et al. [13] introduced a graphical interface named Nugget Browser allowing users to interactively submit subgroup mining queries for discovering interesting patterns dynamically. Zhou et al. [37] revealed states of key internal variables of ML models with interactive visualization to keep users aware what is going on inside a model. More recent work tries to use visualization as an interactive tool to facilitate ML diagnosis. ModelTracker [1] provides an intuitive visualization interface for ML performance analysis and debugging. Chen et al. [10] proposed an interactive visualization tool by combining ten state-of-the-art visualization methods in ML to help users interactively carry out multi-step diagnosis for ML models. Recently, visualization approaches are also proposed to explain complex deep neural networks [14]. For example, saliency maps are used to explain contributions of different points of a data to predictions [6].

Besides visualization, various mathematical approaches are proposed to explain ML models. Robnik-Sikonja et al. [30] explained classification models by evaluating contributions of features to classifications based on the idea that importance of a feature or a group of features in a specific model can be estimated by simulating the lack of knowledge about the values of the feature(s). Besides feature contributions, explanation of individual instance contributions to ML models was investigated to allow users to understand why a classification/prediction is made. For example, Landecker et al. [21] developed an approach of contribution propagation to give per-instance explanations of a network's classifications. Koh et al. [18] used influence functions to evaluate influence of each training data point on predictions.

These approaches explain ML models mostly from an ML expert's perspective, which introduce further complexities to domain users and make users more difficult to understand complex algorithms. Furthermore, these explanations mostly focus on the stage of ML models and pay less attention to the stage of deployment of ML models.

2.2   User Trust in Machine Learning

As the ultimate frontline users of ML-based systems, humans are the key stakeholders and human factors such as user trust are essential in extracting and delivering more sensible and effective insights from data science technologies [12]. From this perspective, Zhou et al. [35,39] argued that communicating user cognitive responses such as trust benefits the evaluation of effectiveness of ML approaches. Therefore, different approaches are investigated to reveal human cognition states such as trust in predictive decision making scenarios [35,37].

Moreover, various researches have been investigated to learn user trust variations in ML. Ye and Johnson [31] experimented with three types of explanations (trace, justification and strategy) for an expert system, and found that justification (defined as showing the rationale behind each step in the decision) was the most effective type of explanation in changing users' attitudes towards the system. Kizilcec [17] proposed that the transparency of algorithm interfaces can promote awareness and foster user trust. It was found that appropriate transparency of algorithms through explanation benefited the user trust. However, too much explanation information on algorithms eroded user trust. Ribeiro et al. [28] explained predictions of classifiers by learning an interpretable model locally around the prediction and visualizing importance of the most relevant features to improve user trust in classifications. Other studies that empirically tested the importance of explanation to users, in various fields such as the health informatics, consistently showed that explanations significantly increase users' confidence and trust [8].

2.3   Physiological Responses in Decision Making

In Human-Computer Interaction (HCI), physiological responses are used to understand an individual's decision making process [39]. For example, GSR refers to how well the skin conducts electricity when an external direct current of constant voltage is applied [11]. It yields continuous signals that are related to activity in the sympathetic branch of the anatomical neural system during tasks. It is well established that skin conductance covaries with the arousal dimension of affect, indexing its intensity. The Iowa Gambling Task (IGT) [3] demonstrated that GSR can be used as a process indicator of affective processes when making decisions. Zhou et al. [39] showed that decision making can be measured with GSR in order to allow users to perceive the quality of their decisions and the level of difficulty involved in making decisions. Therefore, GSR can serve as an objective, non-verbal, non-voluntary indicator and a physiological measure that is relatively free from demand characteristics and reporting biases in

decision making. However, little work has been investigated on the variations of GSR in user trust in a predictive decision making scenario under various conditions such as uncertainty. Furthermore, sympathetic activation has been found to cause changes in heart rate, stroke volume and peripheral cardiovascular resistance [25]. These effects can be sensed by Blood Volume Pulse (BVP), which measures the blood volume in the skin capillary bed in the finger with photoplethysmography (PPG). BVP is often used as an indicator of affective processes and emotional arousal, which play an essential role in rational decision making, learning and cognitive tasks [33]. Zhou et al. [16] showed a set of BVP features for indexing cognitive load.

These previous work motivates us to consider both algorithmic explanations and model performance in the interpretability of ML, aiming to find what explanations and how these explanations affect user trust in ML. We also aim to investigate physiological indicators which may correlate with user trust in predictive decision making. This paper uses the influence of training data points as an example and investigates what influence and how they affect user trust in a predictive decision making scenario.

3   Hypotheses

The following hypotheses are posed in this study:

- H1: The presentation of influence of training data points on predictions will affect the user trust and result in the increase of user trust in predictions;
- H2: The training data points which have the higher influence on predictions will have the higher effect on user trust than those with the lower influence;
- H3: Higher model performance together with the presentation of influence of training data points will result in the higher user trust;
- H4: There are correlations between physiological indicators and user trust under different influence and model performance conditions.

4   Method

In this section, a framework of fact-checking for boosting user trust is firstly presented. A case study is then introduced. After that, the influence of training data points is formulated to understand contributions of training data points to test data predictions. Finally, the fact-checking visualization is proposed to present influence of training data points on test data predictions to users.

4.1   Framework of Fact-Checking for Boosting User Trust

We present a framework of fact-checking for boosting user trust in a predictive decision making scenario (see Fig. 1). In a typical conventional ML pipeline, a training data is used to train an ML model and predictions are made based on the trained model (as shown in the lower unshaded part in Fig. 1). There is

no information on the ML explanation in order to promote the trustworthiness of the prediction. Motivated by the online fact-checking services for strengthening trust [7], an influence-enhanced fact-checking approach is added on the top of the conventional ML pipeline in the proposed framework (as shown in the upper shaded part in Fig. 1) to explain predictions and boost user trust in predictions. Firstly, the influence of all training data points for the prediction of a testing data point is calculated with influence functions as presented in the following subsection. All training data points are then ranked in descending order based on the calculated influence values. Training data points which have the higher influence values (e.g. the top 10 training data points in the ranking) and training data points which have the lower influence values (e.g. the bottom 10 training data points in the ranking) are obtained respectively based on the ranking. These training data points function as facts which are the most similar points to the testing data point and the least similar points to the testing data point respectively. The parallel coordinate based visualization as presented in the following subsection is used to visualize these selected ranked training data points allowing users to compare the facts with the testing data points to boost trust in predictions.
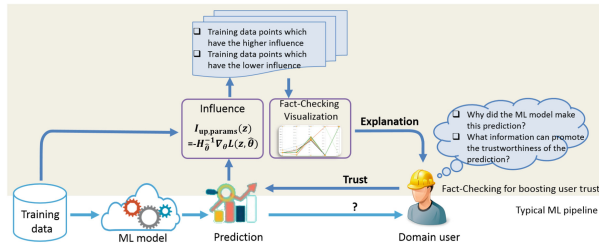

Fig. 1. A framework of fact-checking for boosting user trust.

4.2 Case Study

This paper used water pipe failure prediction as a case study for predictive decision making (replicated in lab environment). Water supply networks constitute one of the most crucial and valuable urban assets. The combination of growing populations and aging pipe networks requires water utilities to develop advanced risk management strategies in order to maintain their distribution systems in a financially viable way [23,40]. Pipes are characterized by different attributes, referred as features, such as laid year, material, diameter size, etc. If pipe failure historical data is provided, future water pipe failure rate is predictable with respect to the inspected length of the water pipe network [23,40]. Such models are used by utility companies for budget planning and pipe maintenance. However, different models with various presentation of influence of training data points and various prediction performance (accuracy) may be achievable resulting in different possible management decisions. The experiment is then set up to determine what influence and model performance may affect the user trust.

4.3   Influence of Training Data Points

Consider a machine learning based prediction problem from an input space $X \in \mathbb{R}^D$ (e.g. water pipe attributes of D dimensions) to an output space $Y \in \{0, 1\}$ (e.g. labels on failures of water pipes), with given training data points $z_1, ..., z_N$, where $z_i = (\boldsymbol{x_i}, y_i) \in X \times Y$, each y_i is a failure observation of each pipe in one year. Based on the data points, a model parameter $\hat{\theta} \in \Theta$ can be learned by minimizing the loss function $\sum_{i=1}^{N} L(z_i, \theta)$. The influence is then defined as how important is a training data point z_{train} to the prediction of a testing data point z_{test}. The influence is then calculated for each pair of (z_{train}, z_{test}).

The intuitive way to get influence is to compare the difference of the prediction results, i.e. y_{test} with and without z_{train} used in the training. The method cannot be scaled up as it requires retraining the model for all training data points for each testing data, which means $N+1$ retraining are needed. This is infeasible for large datasets usually with millions of data points. In this study, the influence function is used to avoid the retraining. It is used to trace a model's prediction through the learning algorithm and back to its training data, thereby identifying training data points most responsible for a given prediction [18]. For any training data point z_{train}, if its weight is to be upweighted by an infinitesimal amount $\epsilon > 0$ from $\frac{1}{N}$, the influence to $\hat{\theta}$ will be quantified as:

$$I(z_{train}) = \frac{d\hat{\theta}_{\epsilon, z_{train}}}{d\epsilon} = -H_{\hat{\theta}}^{-1} \nabla_\theta L(z_{train}, \hat{\theta}) \tag{1}$$

where $H_{\hat{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta^2 L(z_i, \hat{\theta})$ is the Hessian. Then the influence by removing z_{train} can be approximated by $-\frac{1}{N} I(z_{train})$ and chain rule, for z_{test}, it is proportional to a close-form expression:

$$I_{removing}(z_{train}, z_{test}) \propto \nabla_\theta L(z_{test}, \hat{\theta}) I(z_{train}) \tag{2}$$

which can be used in our influence evaluation as we only need the influence ranking of all training data points in the user trust evaluation. The details of these influence functions can be found in [18].

4.4   Fact-Checking Visualization

A training data point in ML usually has multiple features/attributes. The parallel-coordinates is one classical approach to visualize multi-attribute data points. One advantage of this technique is that it can provide an overview of data trend, where each attribute is represented with one axis in parallel coordinates. In this study, we present a visualization approach for presenting multiple data attributes based on parallel coordinates as shown in Fig. 2: each vertical axis represents one data attribute with the sorted descending order, and a polyline connecting points on each vertical axis represents a data point. In this study, each polyline represents one water pipe with various attributes. Various pipe attributes belonging to one pipe are encoded with the same color. Testing pipe

is encoded with red color. The influence of each training pipe on the prediction of a test is encoded with the width of polylines, the wider the polyline, the higher the influence. Such color and line width encoding approach provides an overview of data trend of pipes and their associated attribute details, which can improve the information browsing efficiency. For example, Fig. 2 demonstrates how similar the training pipes are with the testing pipe in red color. If training pipes are considered as facts for predictions, this parallel coordinates based visualization is fact-checking visualization. The pipe attributes visualized in Fig. 2 include pipe size, length, pipe age, failure times during the observation period, and whether it was failed in the checked year (0 means no failure and 1 means failure occured).
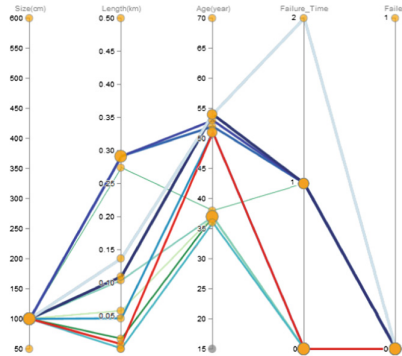

Fig. 2. Fact-checking visualization.

5   Experiment

This section sets up an experiment to examine our hypotheses with the case study of a decision making scenario on water pipe management.

5.1   Experimental Data

Water pipe failure prediction uses historical pipe failure data to predict future failure rate [23]. The historical data contain failure records of water pipes, and various attributes of water pipes, such as laid year, length, diameter size, surrounding soil type, etc. In this study, actual water pipe attributes (features) and the associated historical failure data from a region of a city were used in this experiment. The pipe features used in the experiment include the pipe age, pipe size (diameter), length, and failure times during the observation period. There are 108,745 failure records with 9,062 pipes. 80% of data was used to train the model and the rest was used for the testing. Convolutional neural network (CNN) [20,38] was trained to model the water pipe failures.

In this study, two CNN models (Model 1: 3 hidden layers with number of units of 64, 128, 256, and max iteration number of 1300 respectively; Model 2: 3 hidden layers with number of units of 32, 64, 128, and max iteration number of 700 respectively) were trained using different network settings, resulting in the model accuracy of 90% and 55% respectively. These two model performances were used as high model performance (90%) and low model performance (55%) respectively to find differences of user responses in the experiment. Furthermore, the influence of each training pipes on a prediction was calculated with the use of influence functions introduced in the previous section.

5.2   Task Design

Tasks are designed to investigate how influence of training data points and model performance affect user trust. In this experiment, the training pipes are ranked in the descending order based on their influence values. Based on the ranking, the top 10 (TOP10) and bottom 10 (BOT10) training pipes, which have the highest and lowest influence on predictions respectively, are selected. The fact-checking visualization based on parallel coordinates introduced in the previous section is then used to visualize the TOP10 and BOT10 pipes respectively. Based on the TOP10 and BOT10 pipes' visualization, this experiment divides fact-checking visualization settings for tasks into four categories: (1) *TOP10*, (2) *BOT10*, (3) *TOP10&BOT10* which includes both TOP10 and BOT10 visualizations in tasks, and (4) *Control* which does not include any influence visualization on training pipes. By considering both model performance cases (high and low performance) and fact-checking visualization conditions, we finally got 8 tasks as shown in Table 1. These 8 tasks were conducted two rounds with all same settings except testing pipes used. Two training tasks were also conducted by each participant before formal tasks. In summary, there were 18 tasks conducted (8 tasks × 2 rounds + 2 training tasks = 18 tasks) by each participant.

Table 1. Task setup in the experiment.

		Influence			
		TOP10	BOT10	TOP10&BOT10	Control
Model performance	High	T1	T2	T3	T4
	Low	T5	T6	T7	T8

The decision tasks investigated are: each participant was told that he/she would take the asset management responsibility of a water company. The water company plans to repair pipe failures in the next financial year. He/she was asked to make a decision on whether to replace a testing pipe, using water pipe failure prediction models learned from the historical water pipe failure records. Each task was divided into three stages: at the beginning of each task, participants

were told that a pipe is predicted to fail next year with a prediction accuracy of 90% (High) or 55% (Low); then different fact-checking visualizations based on task settings (see Table 1) are displayed; lastly, participants were asked to make a decision of whether or not to replace the pipe based on the prediction. Participants were told that they were competing against other people to reach the best budget plan in a given time period in order to push them to make their efforts for tasks. The task orders were randomized during the experiment.

5.3   Participants and Data Collection

22 participants were recruited, who are mainly researchers and students with the range of ages from twenties to forties and an average age of 30 years. Of all participants, 5 were females. After each decision making task, participants were asked to rate the trust level of predictions on which decisions were made using a 9-point Likert scale (from 1: least trust, to 9: most trust). Participants were asked to rate how helpful the presentation of influence is for decision making. At the end of each round, participants were also asked to rate the usefulness of influence on helping them more confident in decision making. Besides subjective ratings, skin conductance responses of subjects with GSR sensors and blood volume pulse information with BVP sensors from ProComp Infiniti of Thought Technology Ltd were collected during task time.

6   Analysis of Subjective Ratings

In this study, we aim to understand: (1) the effects of influence on user trust under a given model performance, and (2) the effects of model performance on user trust under a given influence condition respectively. Therefore, for the evaluation of each aims, we first performed Friedman test and then followed it up with post-hoc analysis using Wilcoxon signed-rank tests (with a Bonferroni correction) to analyze differences in participant responses of trust under a fixed condition (e.g. trust changes with different influence types under the fixed high model performance). Trust values were normalized with respect to each subject to minimize individual differences in rating behavior (see Eq. 3):

$$T_i^N = \frac{T_i - T_i^{min}}{T_i^{max} - T_i^{min}} \tag{3}$$

where T_i and T_i^N are the original trust rating and the normalized trusting rating respectively from the user i, T_i^{min} and T_i^{max} are the minimum and maximum of trust ratings respectively from the user i in all of his/her tasks.

6.1   Influence and Trust

Figure 3(a) shows mean normalized trust values over different influence settings under high model performance (error bars represent the 95% confidence interval

of a mean and it is same in other figures). Friedman's test gave statistically significant differences in trust among four influence conditions, $\chi^2(3) = 21.675$, $p = .000$. Then post-hoc Wilcoxon tests (with a Bonferroni correction under a significance level set at $p < .013$) was applied to find pair-wise differences between influence conditions. The adjusted significance alpha level of .013 was calculated by dividing the original alpha of .05 by 4, based on the fact that we had four influence conditions to test.
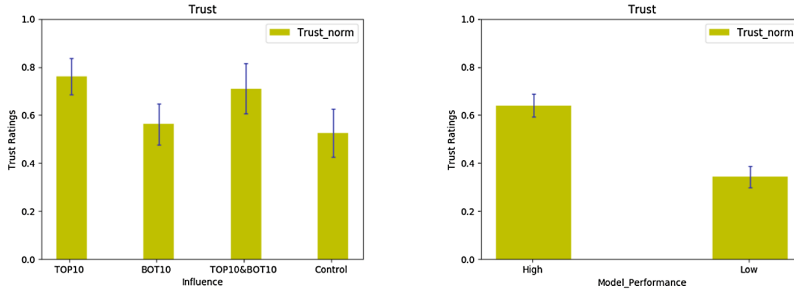
(a) Normalized trust by influence under high model performance.

(b) Normalized trust by different model performance.

Fig. 3. Normalized trust values.

The post-hoc tests found that participants had significantly higher trust in predictions when influences of TOP10 training pipes were presented than those without influence information presentation (Control condition) ($Z = 102.0, p < .001$). Participants also showed significantly higher trust in predictions when influences of both TOP10 and BOT10 were presented than that without influence information presentation (Control condition) ($Z = 120.5, p < .004$). The results suggest that the presentation of influence of training data points on predictions significantly increases the user trust in predictions as we hypothesized (H1). It was also found that participants had significantly higher trust in predictions when influences of TOP10 training pipes were presented than that when influences of BOT10 training pipes were presented ($Z = 61.5, p < .001$). This implies that the training data points having the higher influence on predictions have the higher effect on user trust than that having the lower influence (H2).

However under low model performance, statistically significant differences of trust among different influence conditions have not been found.

These results suggest that the presentation of influence of training data points on predictions significantly increases the user trust in predictions, but only for training data points with higher influence values under the high model performance condition.

6.2   Model Performance and Trust

Figure 3(b) shows mean normalized trust values under two model performance conditions (high and low). A Wilcoxon test found that participants had statistically higher trust in predictions under high model performance than that under low model performance ($Z = 854.5, p = .000$). This result confirms the findings in [32]. We then further drilled down to compare user trust differences over model performance under different influence conditions. Figure 4 shows mean normalized trust values over two model performance conditions (high and low) under different influence settings. It was found that participants showed significantly higher trust under high model performance than that under low model performance over all four influence settings (TOP10: $Z = 11.5, p < .000$; BOT10: $Z = 52.0, p < .000$; TOP10&BOT10: $Z = 77.5, p < .000$; Control: $Z = 77.5, p < .001$). The results suggest that high model performance together with influence information result in the higher user trust in predictions (H3). These findings go on to support the idea that people trust more in predictions with high model performance.
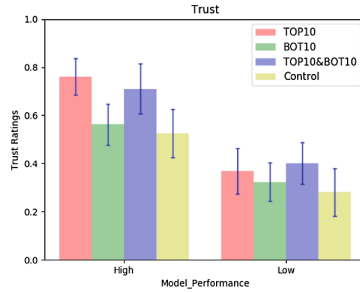

Fig. 4. Normalized trust by model performance under different influences.

7   Physiological Indicators

In this section, GSR and BVP signals are analyzed to investigate their variations under different conditions. The GSR and BVP data analysis process is divided into following steps: (1) signal smoothing, (2) data normalization, (3) feature extraction, and (4) feature significance test.

Similar to the analysis of subjective ratings in the previous section, various GSR and BVP features are also analysed to find: (1) the effects of influence on GSR and BVP features under a given model performance, and (2) the effects of model performance on GSR and BVP features under a given influence condition respectively. Therefore, for the evaluation of each aims, we first performed one-way ANOVA test and then followed it up with post-hoc analysis using t-test (with a Bonferroni correction) to analyze differences in physiological features under a fixed condition.

7.1   Signal Smoothing

The first step of physiological signal analysis is the signal smoothing. For GSR signals, we use convolution filter (similar to a low pass filter) to remove noise. All GSR signals are convoluted to a Hann window function to remove the noise [34]. For BVP signals, we use spectrogram to detect corrupt signals and abnormal outliers, then remove all corrupt data from the dataset [25].

7.2   Normalization

After signal smoothing, we normalize smoothed signal using subject-wise Z-Normalization to omit subjective differences between different signals [34]. The subject-wise normalization means that the mean and variance used in the normalization as in Eq. 4 are from signals of all tasks from each subject.

$$S_N = \frac{S - \mu}{\sigma} \tag{4}$$

where S is the original GSR/BVP value, S_N is the normalized GSR/BVP value, μ and σ are mean and variance respectively of GSR/BVP signals among all tasks conducted by each subject.

7.3   GSR
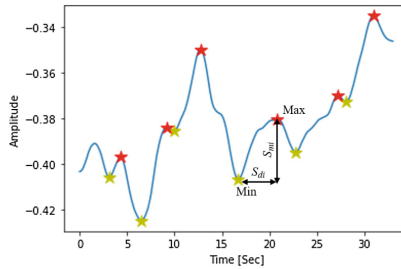
Fig. 5. An example of GSR signals after noise filtering. (Color figure online)

GSR Feature Extractions. In this paper, GSR features are defined based on signal extremas. The extremas are extracted from the normalized GSR signal. Figure 5 shows an example of extremas (red star as local maxima and yellow star as local minima) of GSR. We extracted and analysed both extreme-based and statistical features. All the features are listed in Table 2 [39]. The definition of duration S_{di} and magnitude S_{mi} are shown in Fig. 5. t is the time spent on each task. The estimated area can be regarded as the area of the triangle made by S_d and S_m, which is $S_a = \frac{1}{2}S_d S_m$ [39].

Table 2. GSR features

GSR features	Notes
Mean of GSR μ_g	Summation of all GSR values divided by task time
GSR variance σ_g	Variance of GSR values over task time
Number of peaks S_f	Number of peaks in a GSR signal
Sum of duration S_d	Sum of duration time of all tasks: $S_d = \sum S_{di}$
Sum of magnitude S_m	Sum of magnitude: $S_m = \sum S_{mi}$
Estimated area S_a	Sum of estimated area: $S_a = \sum S_{ai}$
Number of response per second S_{fs}	$S_{fs} = S_f/t$
Duration per second S_{ds}	$S_{ds} = S_d/t$
Magnitude per second S_{ms}	$S_{ms} = S_m/t$
Maximum of duration S_d^{max}	$S_d^{max} = max(S_d)$
Maximum of magnitude S_m^{max}	$S_m^{max} = max(S_m)$
Maximum of estimated area S_a^{max}	$S_a^{max} = max(S_a)$
Average gradient G_a	$G_a = \frac{1}{S_f} \sum S_{mi}/S_{di}$
Maximum gradient G^{max}	$G^{max} = max(S_{mi}/S_{di})$
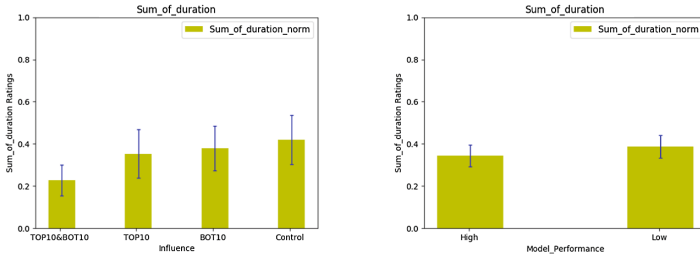
(a) Sum of duration over different influence conditions. (b) Sum of duration over different model performance conditions.

Fig. 6. GSR feature of sum of duration.

GSR Features and Influence. Under the high model performance, one-way ANOVA tests found that there are significant differences in GSR values among different influence conditions for GSR sum of duration S_d ($F(3, 112) = 2.874, p = .039$) (see Fig. 6(a)). The post-hoc t-tests (with a Bonferroni correction under $\alpha = .05/4 = .013$, based on the fact that 4 levels were tested) was used to examine the pair difference between influence conditions in GSR feature of S_d. The post-hoc tests found that participants had significantly lower GSR S_d values when influences of TOP10&BOT10 training pipes were presented than that without influence information presentation (Control condition) ($t = -3.039, p < .004$). Participants also showed relatively lower GSR S_d values when influences of TOP10&BOT10 training pipes were presented than that when BOT10 training pipes were presented. The results suggest that the presentation

of influence of training data points on predictions especially TOP10 training points significantly decreases the GSR S_d values under high model performance.

GSR Features and Model Performance. Figure 6(b) shows mean GSR S_d values over two model performance conditions (high and low). Although there is no significant difference found in GSR S_d values over two model performance conditions, a trend shows that GSR S_d values under high model performance condition is relatively lower than that under low model performance condition. The findings suggest that the high model performance condition has a trend to decrease GSR values.

7.4   BVP

BVP Features. BVP is a periodical signal and associated with three major frequency bands: Very Low Frequency (VLF) (0.00–0.04 Hz), Low Frequency (LF) (0.05–0.15 Hz), and High Frequency (HF) (0.16–0.40 Hz). The LF/HF ratio is calculated by finding the ratio of low frequency energy to high frequency energy in the spectrum. Furthermore, the BVP sensor measures one of physiological changes known as Heart Rate Variability (HRV). HRV is known to be closely related to Respiratory Sinus Arrhythmia (RSA) which can be used as a measurement to quantify the activity of the parasympathetic activity [25,27]. Therefore, both statistical and frequency domain features of BVP are extracted for analysis in this section. Table 3 lists BVP features extracted in this study.

Table 3. BVP features

BVP features	Notes
Mean of BVP μ_b	Summation of all BVP values divided by task time
BVP variance σ_b	Variance of BVP values over task time
Number of peaks S_p	Number of peaks in a BVP signal
BVP peak mean μ_{bp}	Summation of all BVP values divided by number of peaks
BVP peak variance σ_{bp}	Variance of BVP peak values
LF	Power Spectral Density (PSD) for low frequency
HF	Power Spectral Density (PSD) for high frequency
LF HF ratio S_r	$S_r = LF/HF$

BVP Features and Influence. Under the high model performance, one-way ANOVA tests found significant differences in BVP peak mean μ_{bp} values ($F(3, 120) = 4.705$, $p = .004$) and BVP peak variance σ_{bp} values ($F(3, 120) = 4.961, p = .003$) among different influence conditions respectively. Figure 7 shows BVP peak mean values over four influence conditions. The post-hoc t-tests(with a Bonferroni correction as mentioned previously) were conducted to examine the pairwise difference between influence conditions in BVP features. For the BVP peak mean μ_{bp}, it was found that there were statistically

significant lower values in TOP10&BOT10 than that in other three influence conditions TOP10 ($t = -2.921, p = .005$), BOT10 ($t = -3.45, p = .001$), and Control ($t = -3.644, p = .001$) respectively. Similarly, The BVP peak variance σ_{bp} showed significantly lower values in TOP10&BOT10 than that in other three influence conditions TOP10 ($t = -2.947, p = .005$), BOT10 ($t = -3.57, p = .001$), and Control ($t = -3.788, p = .000$) respectively.
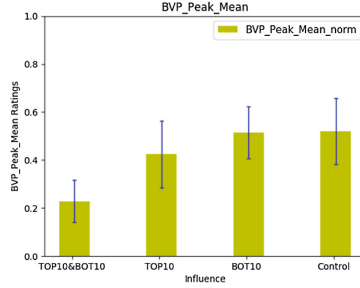
Fig. 7. BVP feature of BVP peak mean over four influence conditions.

These results show that the presentation of influence of training data points on predictions especially both TOP10 and BOT10 training points at the same time significantly decreases BVP values such as μ_{bp} under high model performance. However, all extracted BVP features did not show significant differences over model performance conditions, despite the trend with relative lower GSR values under high model performance related to influence presentations of TOP10 training points.

In summary, we found that the presentation of influence of training data points on predictions especially TOP10 training points significantly decreases both GSR and BVP values such as μ_{bp}, but only under high model performance. Furthermore, a trend shows that both GSR and BVP values are relatively lower under high model performance than that under low model performance. By considering the relations between trust and influence/model performance concluded in the previous section, the findings in this section on GSR and BVP features can be used as indicators of user trust in predictive decision making under different influence and model performance conditions as we expected in H4.

8   Discussions

As discussed in earlier sections, trust is a challenging concept to investigate in machine learning based solutions. This paper intends to study human-machine trust in a specialized predictive decision making scenario. As machines are becoming more intelligent, however in many scenarios, instead of full autonomy, Human-Machine Teaming (HMT) is required, where humans interact with

the intelligent (AI) system to understand why the AI system is suggesting something that the human should do or not do. Therefore, both interaction with and transparency of the system help humans make effective uses of the AI system for trusting decisions.

In the water pipe failure prediction example as mentioned, when pipe management staff want to use an AI tool to make decisions on pipe replacement, they need to be confident that there is a clear rationale for the ML to predict a pipe's future failure, in order to build trust. Therefore, similar to precedent that humans justify actions by analogy, the pipe management staff interact with such kind of AI system to find similar cases (based on pipe features such as material, age, size, and length) to support a planned pipe management protocol. In our approach, the influence values of training pipes on the prediction were used to help users to locate/identify pipes having higher influence values (which may show more similar feature patterns to the testing pipe) or pipes having lower influence values (which may show more dissimilar feature patterns to the testing pipe). These pipes were presented to users with parallel coordinates based visualizations to help users easily get the overall patterns of features of pipes. The interaction with the visualization of pipes functioned as the fact-checking for the prediction to help users understand why a similar or different decision was made, thereby increases the transparency of the system and boosts user trust.

As we have seen that participants showed significantly higher trust when TOP10 visualization was presented. It was also found that participants showed significantly higher trust under high model performance. GSR and BVP features showed correlation to both influence and model performance conditions, suggesting that GSR and BVP features can be used as indicators for trust variations in predictive decision making.

In order to make ML-driven AI applications not only intelligent but also intelligible, the user interface of AI applications needs to allow users to access the most influential facts to predictions by visualizations. Such influence-enhanced fact-checking allows users find similar facts to the testing data point to get the rational behind for the justification of their actions, therefore boosting user trust.

A weakness of this training data based influence interpretation approach is the privacy issue of training data. The proposed approach is not applicable if the training data is sensitive and/or needs to be made private. However, there are still many applications where it is not an issue (such as the water pipe failures).

9   Conclusions and Future Work

This paper investigates the influence enhanced fact-checking for the ML explanation to boost user trust in a predictive decision making scenario. Both influence of training data points on predictions and model performance were examined to find their effects on trust. Physiological features were analysed and showed their correlations to influence and model performance conditions. A user study found that the presentation of influence of training data points on predictions significantly increased the user trust in predictions, but only for training data points

with higher influence values under the high model performance condition, where users were expected to be able to justify their actions with more similar facts to the testing data point. These findings suggested that the access of the most influential facts to predictions by users in the user interface of AI applications would help users get the rational behind their actions and therefore benefit the user trust in predictions.

Our future work will focus on the setup of ML models to automatically predict user trust in decision making based on physiological features, which contributes to the ultimate goal of intelligent user interface of AI applications.

References

1. Amershi, S., Chickering, M., Drucker, S.M., Lee, B., Simard, P., Suh, J.: ModelTracker: redesigning performance analysis tools for machine learning. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 337–346 (2015)
2. Ankerst, M., Elsen, C., Ester, M., Kriegel, H.P.: Visual classification: an interactive approach to decision tree construction. In: Proceedings of KDD 1999, pp. 392–396 (1999)
3. Bechara, A., Damasio, H., Damasio, A.R., Lee, G.P.: Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. J. Neurosci. **19**, 5473–5481 (1999)
4. Becker, B., Kohavi, R., Sommerfield, D.: Visualizing the simple Bayesian classifier. In: Fayyad, U., Grinstein, G.G., Wierse, A. (eds.) Information visualization in data mining and knowledge discovery, pp. 237–249 (2002)
5. Biran, O., Cotton, C.: Explanation and justification in machine learning: a survey. In: Proceedings of the 2017 IJCAI Explainable AI Workshop, pp. 8–13 (2017)
6. Brahimi, M., Arsenovic, M., Laraba, S., Sladojevic, S., Boukhalfa, K., Moussaoui, A.: Deep learning for plant diseases: detection and saliency map visualisation. In: Zhou, J., Chen, F. (eds.) Human and Machine Learning. HIS, pp. 93–117. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-90403-0_6
7. Brandtzaeg, P.B., Følstad, A.: Trust and distrust in online fact-checking services. Commun. ACM **60**(9), 65–71 (2017)
8. Calero Valdez, A., Ziefle, M., Verbert, K., Felfernig, A., Holzinger, A.: Recommender systems for health informatics: state-of-the-art and future perspectives. In: Holzinger, A. (ed.) Machine Learning for Health Informatics. LNCS (LNAI), vol. 9605, pp. 391–414. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50478-0_20
9. Caragea, D., Cook, D., Honavar, V.G.: Gaining insights into support vector machine pattern classifiers using projection-based tour methods. In: Proceedings of KDD 2001, pp. 251–256 (2001)
10. Chen, D., Bellamy, R.K.E., Malkin, P.K., Erickson, T.: Diagnostic visualization for non-expert machine learning practitioners: a design study. In: 2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pp. 87–95, September 2016

11. Figner, B., Murphy, R.O.: Using skin conductance in judgment and decision making research. In: A Handbook of Process Tracing Methods for Decision Research: A Critical Review And User's Guide, pp. 163–184 (2010)
12. Fisher, D., DeLine, R., Czerwinski, M., Drucker, S.: Interactions with big data analytics. Interactions **19**(3), 50–59 (2012)
13. Guo, Z., Ward, M.O., Rundensteiner, E.A.: Nugget browser: visual subgroup mining and statistical significance discovery in multivariate datasets. In: Proceedings of the 15th International Conference on Information Visualisation, pp. 267–275 (2011)
14. Hartono, P.: A transparent cancer classifier. Health Inform. J. (2018)
15. Ilyas, A., Engstrom, L., Athalye, A., Lin, J.: Black-box adversarial attacks with limited queries and information. In: Dy, J., Krause, A. (eds.) Machine Learning. PMLR, vol. 80, pp. 2142–2151. Stockholmsmässan, Stockholm (2018)
16. Zhou, J., Arshad, S.Z., Luo, S., Yu, K., Berkovsky, S., Chen, F.: Indexing cognitive load using blood volume pulse features. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI EA 2017, May 2017
17. Kizilcec, R.F.: How much information?: Effects of transparency on trust in an algorithmic interface. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 2390–2395 (2016)
18. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, pp. 1885–1894, 6–11 August 2017
19. Kriplean, T., Bonnar, C., Borning, A., Kinney, B., Gill, B.: Integrating on-demand fact-checking with public dialogue. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and & Social Computing, CSCW 2014, pp. 1188–1199 (2014)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105 (2012)
21. Landecker, W., Thomure, M.D., Bettencourt, L.M.A., Mitchell, M., Kenyon, G.T., Brumby, S.P.: Interpreting individual classifications of hierarchical networks. In: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 32–38, April 2013
22. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors **46**(1), 50–80 (2004)
23. Li, Z., et al.: Water pipe condition assessment: a hierarchical beta process approach for sparse incident data. Mach. Learn. **95**(1), 11–26 (2014)
24. Lipton, Z.C.: The mythos of model interpretability. In: Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY, USA (2016)
25. Luo, S., Zhou, J., Duh, H.B.L., Chen, F.: BVP feature signal analysis for intelligent user interface. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA 2017, pp. 1861–1868 (2017)
26. Mannarswamy, S., Roy, S.: Evolving AI from research to real life - some challenges and suggestions. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pp. 5172–5179. International Joint Conferences on Artificial Intelligence Organization, July 2018
27. Nilsson, M., Funk, P.: A case-based classification of respiratory sinus arrhythmia. In: Funk, P., González Calero, P.A. (eds.) ECCBR 2004. LNCS (LNAI), vol. 3155, pp. 673–685. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28631-8_49

28. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. arXiv:1602.04938 [cs, stat], February 2016
29. Richardson, A., Rosenfeld, A.: A survey of interpretability and explainability in human-agent systems. In: Proceedings of IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence (XAI), pp. 137–143 (2018)
30. Robnik-Sikonja, M., Kononenko, I., Strumbelj, E.: Quality of classification explanations with PRBF. Neurocomputing **96**, 37–46 (2012)
31. Ye, L.R., Johnson, P.E.: The impact of explanation facilities on user acceptance of expert systems advice. MIS Q. **19**(2), 157–172 (1995)
32. Yin, M., Vaughan, J.W., Wallach, H.: Does stated accuracy affect trust in machine learning algorithms? In: Proceedings of ICML2018 Workshop on Human Interpretability in Machine Learning (WHI 2018), July 2018
33. Zhai, J., Barreto, A., Chin, C., Li, C.: Realization of stress detection using psychophysiological signals for improvement of human-computer interactions. In: Proceedings of IEEE SoutheastCon 2005, pp. 415–420 (2005)
34. Zhou, J., Arshad, S.Z., Wang, X., Li, Z., Feng, D., Chen, F.: End-user development for interactive data analytics: uncertainty, correlation and user confidence. IEEE Trans. Affect. Comput. **9**(3), 383–395 (2018)
35. Zhou, J., Bridon, C., Chen, F., Khawaji, A., Wang, Y.: Be informed and be involved: effects of uncertainty and correlation on user confidence in decision making. In: Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems (CHI2015) Works-in-Progress, Korea (2015)
36. Zhou, J., Chen, F. (eds.): Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-90403-0
37. Zhou, J., Khawaja, M.A., Li, Z., Sun, J., Wang, Y., Chen, F.: Making machine learning useable by revealing internal states update - a transparent approach. Int. J. Comput. Sci. Eng. **13**(4), 378–389 (2016)
38. Zhou, J., Li, Z., Zhi, W., Liang, B., Moses, D., Dawes, L.: Using convolutional neural networks and transfer learning for bone age classification. In: 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA 2017), pp. 1–6 (2017)
39. Zhou, J., et al.: Measurable decision making with GSR and pupillary analysis for intelligent user interface. ACM Trans. Comput.-Hum. Interact. **21**(6), 33 (2015)
40. Zhou, J., Sun, J., Wang, Y., Chen, F.: Wrapping practical problems into a machine learning framework: using water pipe failure prediction as a case study. Int. J. Intell. Syst. Technol. Appl. **16**(3), 191–207 (2017)