# Real-Time Sensing of Trust in Human-Machine Interactions ⋆

**Wan-Lin Hu** * **Kumar Akash** * **Neera Jain** * **Tahira Reid** *

* *School of Mechanical Engineering, Purdue University, West Lafayette, IN 47906 USA (e-mail: hu188@purdue.edu, kakash@purdue.edu, neerajain@purdue.edu, tahira@purdue.edu).*

**Abstract:** Human trust in automation plays an important role in successful interactions between humans and machines. To design intelligent machines that can respond to changes in human trust, real-time sensing of trust level is needed. In this paper, we describe an empirical trust sensor model that maps psychophysiological measurements to human trust level. The use of psychophysiological measurements is motivated by their ability to capture a human's response in real time. An exhaustive feature set is considered, and a rigorous statistical approach is used to determine a reduced set of ten features. Multiple classification methods are considered for mapping the reduced feature set to the categorical trust level. The results show that psychophysiological measurements can be used to sense trust in real-time. Moreover, a mean accuracy of 71.57% is achieved using a combination of classifiers to model trust level in each human subject. Future work will consider the effect of human demographics on feature selection and modeling.

*Keywords:* human-machine interface, modeling, real-time, categorical data, classifiers, discriminant analysis, human brain, intelligent machines, physiological models

## 1. INTRODUCTION

*Motivation and Problem Definition:* Advances in sensing, communication, and control systems have spurred the development of a number of *smart* systems and services. Increasing levels of automation have resulted in humans being displaced as the primary decision-maker in roles such as power plant operators and aircraft pilots (Jian et al., 2000). Additionally, in what are broadly being called Human-Agent Collectives, we expect to see a growing need for cooperation between humans and machines in a variety of situations, including disaster relief (Jennings et al., 2014; Sadrfaridpour et al., 2016). It is well established that human trust in automation is central to successful interactions between humans and machines (Yagoda and Gillan, 2012; Lee and See, 2004; Sheridan and Parasuraman, 2005). Here, *machine* refers broadly to any automated system, such as an autonomous robot or a process control system in a power plant. Therefore, we are interested in using feedback control principles to design machines that are capable of *responding to changes in human trust level in real-time*. However, in order to do this, we require a sensor for measuring human trust level *online*.

Trust itself can be classified into three categories: dispositional, situational, and learned (Hoff and Bashir, 2015). Dispositional trust refers to the component dependent on demographics (e.g. gender, culture) whereas situational and learned trust depends on time-varying factors such as task difficulty, self-confidence, and experience. Therefore, situational and learned trust factors influence *real-time human decision-making during interactions with automated systems*. Researchers have attempted to predict human trust using dynamic models that rely on the experience and/or self-reported behavior of humans (Lee and Moray, 1992; Jonker and Treur, 1999). However, it is not practical to use human self-reported behavior as a feedback control variable. An alternative is the use of psychophysiological signals to *sense* trust level (Riedl and Javor, 2012). While these measurements have been correlated to human trust level, they have not been studied in the context of real-time trust sensing.

*Background on Psychophysiological Measurements and Trust:* There are several psychophysiological measurements that have been studied in the context of human trust. We focus here on electroencephalography (EEG) and galvanic skin response (GSR). EEG is an electrophysiological measurement technique that captures the cortical activity of the brain (Handy, 2005), and a powerful technique to observe brain activity in response to a specific event is through an event-related potential (ERP). An ERP is determined by averaging repeated responses over many trials to eliminate random brain activity (Handy, 2005). GSR is a classical psychophysiological signal that captures arousal based upon the conductivity of the surface of the skin. It has been used in polygraph tests for many decades (Grubin and Madsen, 2005).

Some researchers have studied trust via EEG, especially with ERPs. Boudreau et al. (2008) found a difference in peak amplitudes of ERP components in human subjects

⋆ This material is based upon work supported by the National Science Foundation under Award No. 1548616. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

while they participated in a coin toss experiment that stimulated trust and distrust. Long et al. (2012) further studied ERP waveforms with feedback stimuli based on a modified form of the coin toss experiment conducted by Boudreau et al. (2008). The decision-making in the 'trust game' (Ma et al., 2015) has also been used to examine human-human trust level. Finally, researchers have examined GSR in correlation with human trust level. Khawaji et al. (2015) found that the average of GSR values, and the average of peaks of GSR values, are significantly affected by both trust and cognitive load in the text-chat environment.

*Gaps in Literature:* Although ERPs could show how the brain functionally responds to a stimulus, they are event triggered. It is difficult to identify triggers during the course of an actual human-machine interaction thereby rendering ERPs impractical for real-time trust level sensing. In addition, the use of GSR for measuring trust has not been explored. A fundamental gap remains in determining a static mathematical model that maps psychophysiological signals to human trust level and that is suitable for real-time sensing.

*Contribution:* In this paper we present a human trust sensor model based upon real-time psychophysiological measurements, primarily GSR and EEG. The model is based upon data collected through a human subject study and the use of classification algorithms to map continuous data to a categorical trust level. The proposed methodology for real-time sensing of human trust level will enable machine algorithm design aimed at improving interactions between humans and machines.

*Outline:* This paper is organized as follows. Section 2 introduces the experimental procedure and data acquisition. The methodology for data analysis is described in Section 3. The sensor modeling and classification results are presented and discussed in Section 4, followed by concluding statements in Section 5.

## 2. HUMAN SUBJECT STUDY

Prior investigation of human trust with respect to psychophysiological response has relied on experiments that do not mimic realistic human-machine interaction (HMI) scenarios (Boudreau et al., 2008; Long et al., 2012). We believe that the use of an experiment in a simple HMI context will result in trust models that are more broadly applicable. Thus we propose the following experiment that elicits human trust dynamics with respect to machines.

*Participants:* Thirty-one adults (20 males) from West Lafayette, Indiana (USA), aged 18-43 years participated in our study. All participants were healthy and one was left-handed. The group of participants were diverse with respect to their age, gender, major, and cultural background (i.e. nationality). The compensation was $15 per hour for their participation and each participant signed the informed consent form. The Institutional Review Board at Purdue University approved the study.

*Stimuli and Procedures:* When a participant came to the laboratory, we asked them to respond to a scenario in which they would be driving a car equipped with an image processing sensor. The algorithm used in the sensor would detect obstacles on the road in front of the car and the participant would need to repeatedly evaluate the algorithm report. We specifically informed the participant that the algorithm for image processing was in beta testing and that they would need to make their judgment of trust or distrust based on their experience with the algorithm.

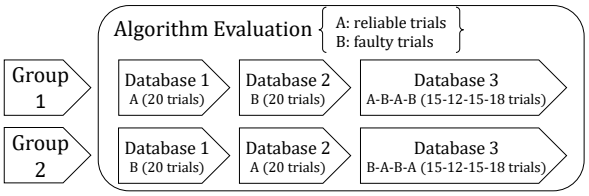



Fig. 1. Participants were randomly assigned to one of the two groups. The ordering of the three experimental sections (databases), composed of reliable and faulty cases, were counterbalanced across groups.

There were two stimuli (*obstacle detected* and *clear road*). Both stimuli had a 50% probability of occurrence. Participants had the option to choose 'trust' or 'distrust' after which they received feedback of 'correct' or 'incorrect'. The trials were divided into two categories: reliable and faulty. In reliable trials, the algorithm correctly identified the road condition, which was in fact the stimuli. From the participant's perspective, this meant that choosing 'trust' would be marked as correct and choosing 'distrust' would be marked as incorrect. For the faulty trials, there was a 50% probability that the algorithm incorrectly identified the road condition.

Each participant completed 100 trials, along with four practice trials in the beginning of the study. The trials were divided into three phases, called databases in the study, as shown in Fig. 1. In database 3, the accuracy of the algorithm was switched between reliable and faculty according to a pseudo-random binary sequence (PRBS) in order to excite all possible dynamics of the participant's trust response. Figure 2 shows the sequence of events in a single trial. We validated the experimental design by collecting responses from 209 online participants (112 and 97 in groups 1 and 2, respectively) using Amazon Mechanical Turk (Amazon, 2005). The experiment elicited expected trust responses based on the aggregated data as shown in Fig. 3.

*EEG Recording and Pre-processing:* EEG, sampled at 256 Hz, was recorded from 9 scalp sites (Fz, Cz, POz, F3, F4, C3, C4, P3, and P4 based on the 10-20 system) using the B-Alert X10 EEG headset (Advanced Brain Monitoring, CA, USA) via iMotions (iMotions, Inc., MA, USA). All EEG channels were referenced to the mean of the left and right mastoids. The surface of the scalp and the mastoids were cleaned with 70% isopropyl alcohol wipes. Conductive electrode cream (Kustomer Kinetics, CA, USA) was then applied to each electrode including the reference. The contact impedance between electrodes and skin was kept to a value less than 40 k$\Omega$.

Automatic decontaminated signals provided by the EEG system were used for model training and validation; that is to say, effects from electromyography, electrooculography, spikes, saturations, and excursions were minimized.

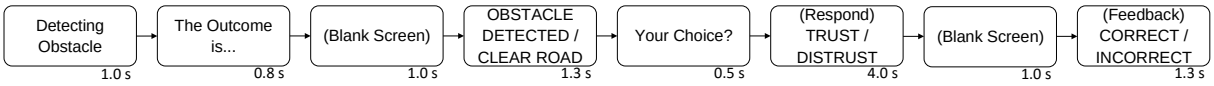
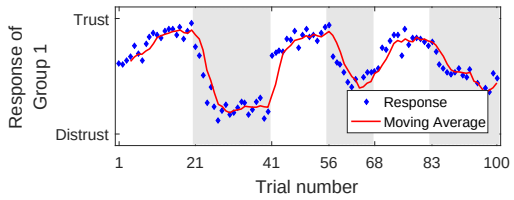
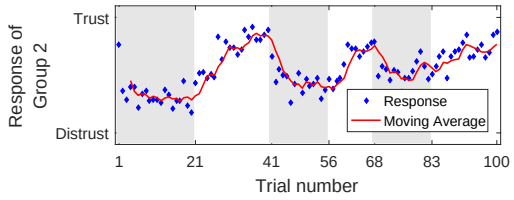



Fig. 2. Sequence of events in a single trial. The time length marked on the bottom right corner indicates the time interval the information was displayed on the screen.




(a)




(b)

Fig. 3. The averaged response from online participants. Faulty trials are highlighted in gray. Participants showed a high trust level in reliable trials and a low trust level in faulty trials regardless of the group they were in.

Before proceeding further, we analyzed the distribution of spectral components of EEG data manually for the entire duration of the experiment for each participant. We eliminated the entire data of participants who had anomalous EEG spectrum, possibly due to bad channels or dislocation of the EEG headset. We removed eight participants' data after pre-processing. EEG measured from F3 and F4 was excluded from the data analysis because it was contaminated with eye movement and blinking (Berka et al., 2007).

*GSR Recording and Pre-processing:* GSR was recorded from the proximal phalanges of the index and the middle fingers of the non-dominant hand (i.e. on the left hand for 30 out of 31 participants) via the Shimmer3 GSR+ Unit (Shimmer, MA, USA). Locations for attaching Ag/AgCl electrodes were prepared with 70% isopropyl alcohol. The participants were asked to keep their hand steady on the desk to minimize the influence of movement on the signals. GSR was sampled at 52 Hz, and downsampled three times. We also applied an adaptive filter to remove noise from the signal.

## 3. DATA ANALYSIS

In order to map continuous EEG and GSR signals to discrete events associated with the human participants' responses, we extracted an *exhaustive set* of potential input variables from the continuous data for each stimuli. We then reduced the dimension of this variable set, from here onwards to be referred to as the feature set, to include only the statistically significant variables of trust.

### *3.1 Feature Extraction*

We divided the complete data set into 100 intervals (epochs), each starting at the instant the stimulus was presented on the screen to the human participant. The epoch length was chosen as the median response time for each of the 23 participants where response time was defined as the time interval between the stimuli and the response. Therefore, each epoch captured the psychophysiological changes of each participant during their response to the stimulus.

*EEG:* We extracted both frequency and time domain features from each epoch. For frequency domain features, we calculated power spectral densities (PSDs) for five spectral bands, namely theta (4 Hz - 7.5 Hz), alpha (7.5 Hz - 12.5 Hz), low-beta (12.5 Hz - 16.5 Hz), mid-beta (16.5 Hz - 20.5 Hz), and high-beta (20.5 Hz - 30 Hz) for seven of the nine channels. This introduced 35 ($5 \times 7$) potential input variables for sensing the trust level. Regarding the time domain features, we included maximum, minimum, mean, median, mean frequency, median frequency, root-mean-square, variance, kurtosis, and peak-to-peak values of each epoch, thus introducing 70 ($10 \times 7$) more potential input variables.

*GSR:* GSR is usually a superposition of the phasic (fast-changing) and tonic (slow-changing) components of the skin response. To ensure a more robust analysis, we used Ledalab and applied Continuous Decomposition Analysis to separate GSR data into continuous signals of tonic and phasic activity (Benedek and Kaernbach, 2010). This decomposition allows the GSR data to be useful in situations with high phasic activity and allows higher flexibility in the analysis. Features including maximum deflection in net signal, maximum phasic component, and net phasic component, were extracted for each epoch. Thus we introduced three more potential input variables of trust.

### *3.2 Feature Selection*

In addition to EEG and GSR features, we selected response time of the participants in each trial as one of the potential input variables to sense trust level. This resulted in 109 ($35+70+3+1$) potential input variables. In order to avoid "the curse of dimensionality" (Lotte et al., 2007), the 109 features were reduced to a smaller feature set as described below.

From Fig. 3, we observed that after a change in the reliability of the sensor, it took an average of five trials for the participant to adapt to the reliability level of the scenario. To ensure the quality of the features, we excluded the data points from this transition region (i.e. the first five trials) and included the data from the last fifteen trials of databases 1 and 2 (the 6-20 and 26-40 trials of the full study). We labeled all data points (potential input variables) based on the test scenario (reliable or

Table 1. Features to be used as input variables

| | Feature | Measurement | Domain |
|---|---|---|---|
| 1 | Net Phasic Component | GSR | Time |
| 2 | Maximum Phasic Component | GSR | Time |
| 3 | High Beta Band - P4 | EEG | Frequency |
| 4 | High Beta Band - POz | EEG | Frequency |
| 5 | High Beta Band - C4 | EEG | Frequency |
| 6 | Mid Beta Band - C3 | EEG | Frequency |
| 7 | Mean Frequency - C3 | EEG | Time |
| 8 | Mean Frequency - C4 | EEG | Time |
| 9 | Mean Frequency - P4 | EEG | Time |
| 10 | Response Time | Behavior | Time |

faulty), which represented the average trust level of the participants. We then selected a subset of features using the *Scalar Feature Selection* technique (Theodoridis and Koutroumbas, 2006). We treated each of the potential input variables as individual features and then used the *Fisher Discriminant Ratio* ($FDR$),

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \quad , \tag{1}$$

as the class separability criterion, where $\mu_1$, $\mu_2$ are the mean values and $\sigma_1$, $\sigma_2$ are the standard variations of class 1 (trust) and 2 (distrust), respectively. We calculated $FDR$ for each feature and for each individual participant. Then, we calculated the class separability score for each feature, $C(k)$, by averaging the $FDR$ values of each feature across participants. We then ranked the features in descending order of values of $C(k)$. This approach did not take into account existing correlations between features because we were treating features individually. In order to incorporate correlation information, we used the following technique. Let $x_{nk}$, $n = 1, 2, ..., N$ and $k = 1, 2, ..., 109$ be the $k$th feature of the $n$th vector ($N = 30$, since we used 30 trials for feature selection). We calculated the cross-correlation coefficient ($\rho_{ij}$) between any two features as

$$\rho_{ij} = \frac{\Sigma_{n=1}^{N} x_{ni} x_{nj}}{\sqrt{\Sigma_{n=1}^{N} x_{ni}^2 \Sigma_{n=1}^{N} x_{nj}^2}} \quad . \tag{2}$$

Lotte et al. (2007) stated that the sample size of the data should be at minimum five to ten times the number of features. With 100 data points (trials) for each participant, we selected ten input variables for classification. We followed these two steps to select an optimal subset of features: (1) select the feature with the best $C$ value and define this feature as $x_{i_1}$; (2) select $x_{i_k}$, $k = 2, 3, ..., 10$ such that

$$i_k = arg\max_j \left\{ \alpha_1 C(j) - \frac{\alpha_2}{k-1} \Sigma_{r=1}^{k-1} |\rho_{i_1 j}| \right\} \tag{3}$$

for $j \neq i_r$, $r = 1, 2, ..., k-1$, $\alpha_1 = 0.7$, and $\alpha_2 = 0.3$. The relative values of $\alpha_1$ and $\alpha_2$ were chosen based on the collinearity of the full feature set; a larger $\alpha_2$ is required when features are highly correlated to each other. By using this method we ensured that subsequent feature selection would take into account the class separability measure $C$ as well as the average correlation of that feature with the already chosen features. The final set of ten features is listed in Table 1.

### *3.3 Discussion*

Among the frequency domain EEG features, high-beta band measurements from the right hemisphere and the parietal lobe (C4, P4, POz) responded most strongly to the discrepancy between reliable and faulty stimuli. High-beta band is positively related to anxiety, activation/excitation, and increased vigilance (Ray and Cole, 1985; Knyazev et al., 2002) while the right hemisphere of the brain is dominant for attention (Heilman and Van Den Abell, 1980). These two features are likely to be significant when a human's trust level in an automated system is low. Mid-beta band is another important feature that has been shown to be related to emotional states in the literature (Isotani et al., 2001) and may represent the emotional component of human trust.

Three time-domain EEG features were also found to be significant; this is consistent with literature in which brain activity is evident in the dynamic variation of measured EEG signals (Lotte et al., 2007). It is noteworthy that both mid-beta band and mean frequency at C3 were found to be correlated with trust. This finding agrees with an ERP study that showed that the N2 component of the ERP at C3 was significantly affected by anxiety (Righi et al., 2009).

GSR showed a promising capability of discriminating the level of trust. Both net phasic component and maximum phasic activity are significant predictors. Khawaji et al. (2015) investigated a similar result with respect to the net phasic component (which corresponds to average peak values), but trust and cognitive load were coupled in their research. In addition, their method depended on the mean GSR which is dominated by the slow-changing (with respect to time) tonic activity and cannot reflect rapid changes in trust level.

Finally, we found that the response time was also a significant feature of trust. This is consistent with findings of Boudreau et al. (2008). They observed that participants responded relatively faster in the common interests (i.e. trust) condition than in the conflicting interests (i.e. distrust) condition in their study.

## 4. MODELING AND VALIDATION

In order to derive a sensor model that maps the ten continuous input variables (table 1) to a categorical output variable (i.e. trust or distrust), we required the use of either regression or classification algorithms. Similar problems have been considered in the design of Brain-Computer Interfaces (BCIs) in which an interface is created to enable a computer or an electronic device to understand a human's command through brain activity. Feature selection and classification algorithms are the most popular approach in BCI design as they typically provide higher accuracy (Mcfarland et al., 2006) compared to regression. Thus we aimed to use classifiers to derive a trust sensor model.

### *4.1 Classifier Training*

We implemented five types of classifiers: Linear Discriminant Analys (LDA), linear Support Vector Machine (SVM), logistic regression, quadratic discriminant, and

Weighted k Nearest Neighbors (kNN) (Theodoridis and Koutroumbas, 2006) using the Statistics and Machine Learning Toolbox in MATLAB R2016a (The MathWorks, Inc., USA). These static discriminant classifiers, which vary in properties including linearity, stability, robustness to high dimensionality, and regularization, have been successfully applied in the design of BCI (Lotte et al., 2007). Combining several classifiers demonstrated some advantages in reducing the classification error (Pfurtscheller et al., 1993; Rakotomamonjy et al., 2005). Therefore, we also used *Voting* (Lotte et al., 2007), a classifier combining strategy, with the above mentioned classifiers.

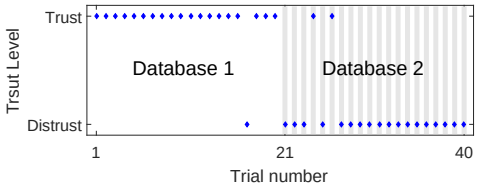
From Fig. 3, it appeared that the trust level of participants did not reach steady state in database 3, as it did in databases 1 and 2. Therefore we trained and validated each of the classifiers using two sets of input feature vectors for each individual participant. The first set consisted of trials from the first two databases (i.e. the first 40 trials). The second set consisted of all 100 trials of the experiment.

### *4.2 Model Validation*

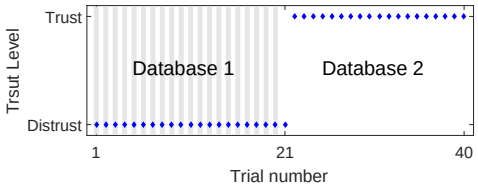
Given the limited size of our dataset, it was not feasible to divide the data into distinct training and testing sets. Therefore, we used a 5-fold cross-validation technique to find the accuracy of each classifier (Hastie et al., 2009). This validation technique randomly divides the data into five sets, and calculates the average test error over all five sets. The test error for each set is evaluated with the model trained from the other four sets. We performed 1000 iterations with different divisions of sets to evaluate the accuracy of each classifier. Tables 2 and 3 show the comparison for mean, maximum, and minimum values of the accuracy of each classifier across participants for the first 40 and then all 100 trials, respectively. The 95% confidence interval (CI) on the mean accuracy is reasonably narrow, which implies the classifier is *robust to the selection of a training data set* for the classifier.

During the first 40 trials, the mean accuracy was $71.57 \pm 0.05\%$. For some individual participants, the mean accuracy increased to $97.23 \pm 0.05\%$. This performance is notable because 10 of the 40 trials (25%) represented a transition in the behavior of each participant and were therefore less predictable. Figures 4(a) and 4(b) are examples of good predictions (i.e. 92.50% and 97.50% accuracy) for group 1 and 2, respectively. Figure 4(c) shows a transition state at the beginning of database 2; it took five trials for this participant to establish a new trust level, and the prediction accuracy was 80%. The classification accuracy is low for some participants as shown in Figure 4(d) (52.50% accurate). The ten features selected for the trust model presented here may not be the best features for all participants due to variations in demographics which affect dispositional trust (e.g. Riedl and Javor (2012)). Future work will address this variation in the model structure.

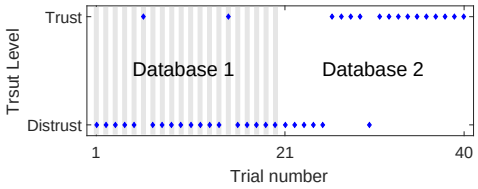
The mean accuracy with voting when all 100 trials were used for classifier training was $60.72 \pm 0.04\%$ while the maximum was $73.33 \pm 0.13\%$ accurate. As shown in Fig. 3, the trust level of the participants does not reach steady state and mostly stays in transition during the PRBS perturbation in the last 60 trials of the study. The binary classifier that we established may not fit this scenario well,



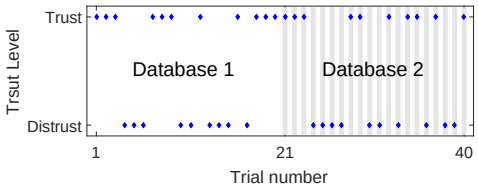

(a) Participant 25 (group 1)




(b) Participant 14 (group 2)




(c) Participant 18 (group 2)




(d) Participant 34 (group 1)

Fig. 4. Example classification results of the first 40 trials based on the voting classifier. Faulty trials are highlighted in gray. Blue dots are the sensed trust level; trust in reliable trials and distrust in faulty trials are considered accurate sensing.

so the classification accuracy is relatively low. Future work will consider a slower perturbation during database 3 in order to allow the participants' trust levels to reach steady state; this will result in larger data training sets. Another potential avenue for exploration would be to consider the transition region as an intermediary trust level.

## 5. CONCLUSION

To achieve more symbiotic interactions between humans and machines, machines must be able to adapt and respond to the trust levels of humans. A critical first step is the design of a real-time trust sensor. In this paper, we described an empirical trust sensor model that maps psychophysiological measurements to human trust level. An exhaustive feature set was considered, and a rigorous statistical approach was used to downselect the set to the best ten features. We then used five different classification methods to derive multiple models between the ten features and the categorical trust level. The results show that psychophysiological measurements can be used to sense trust in real-time. Using the Voting classifier as a model for trust level in each human subject, a mean accuracy of 71.57% was achieved. However, the results also showed that the chosen set of features is not necessarily suitable

Table 2. Accuracy (%) of the classifiers with first 40 trials as input vector with 95% CI

| **Classifier** | Linear Discriminant | Linear SVM | Logistic Regression | Quadratic Discriminant | Weighted KNN | **Voting** |
|---|---|---|---|---|---|---|
| **Mean** | $67.60 \pm 0.05$ | $69.37 \pm 0.06$ | $70.86 \pm 0.07$ | $73.68 \pm 0.05$ | $67.36 \pm 0.05$ | $\mathbf{71.57 \pm 0.05}$ |
| **Max** | $95.34 \pm 0.14$ | $97.24 \pm 0.05$ | $99.36 \pm 0.08$ | $98.35 \pm 0.07$ | $96.88 \pm 0.07$ | $\mathbf{97.23 \pm 0.05}$ |
| **Min** | $41.40 \pm 0.34$ | $39.41 \pm 0.31$ | $37.70 \pm 0.38$ | $45.70 \pm 0.33$ | $41.16 \pm 0.30$ | $\mathbf{42.61 \pm 0.34}$ |

Table 3. Accuracy (%) of the classifiers with all 100 trials as input vector with 95% CI

| **Classifier** | Linear Discriminant | Linear SVM | Logistic Regression | Quadratic Discriminant | Weighted KNN | **Voting** |
|---|---|---|---|---|---|---|
| **Mean** | $57.57 \pm 0.03$ | $59.28 \pm 0.04$ | $58.49 \pm 0.04$ | $59.92 \pm 0.03$ | $56.86 \pm 0.04$ | $\mathbf{60.72 \pm 0.04}$ |
| **Max** | $69.27 \pm 0.13$ | $72.56 \pm 0.12$ | $70.26 \pm 0.10$ | $69.89 \pm 0.07$ | $71.80 \pm 0.11$ | $\mathbf{73.33 \pm 0.13}$ |
| **Min** | $43.62 \pm 0.19$ | $43.27 \pm 0.19$ | $46.33 \pm 0.17$ | $43.30 \pm 0.20$ | $45.39 \pm 0.15$ | $\mathbf{46.61 \pm 0.21}$ |

for all humans. This may be related to dispositional trust features not considered here and will be the subject of future work.

## REFERENCES


Amazon (2005). Amazon mechanical turk. [ONLINE] Available at: https://www.mturk.com/. [Accessed 20 February 2016].

Benedek, M. and Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1), 80–91.

Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., Olmstead, R.E., Tremoulet, P.D., and Craven, P.L. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, 78(Supplement 1), B231–B244.

Boudreau, C., McCubbins, M.D., and Coulson, S. (2008). Knowing when to trust others: An ERP study of decision making after receiving information from unknown people. *Social Cognitive and Affective Neuroscience*, 4(1), 23–34.

Grubin, D. and Madsen, L. (2005). Lie detection and the polygraph: A historical review. *Journal of Forensic Psychiatry & Psychology*, 16(2), 357–369.

Handy, T.C. (2005). *Event-related potentials: A methods handbook*. MIT press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer-Verlag New York.

Heilman, K.M. and Van Den Abell, T. (1980). Right hemisphere dominance for attention The mechanism underlying hemispheric asymmetries of inattention (neglect). *Neurology*, 30(3), 327–327.

Hoff, K.A. and Bashir, M. (2015). Trust in automation integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(3), 407–434.

Isotani, T., Tanaka, H., Lehmann, D., Pascual-Marqui, R.D., Kochi, K., Saito, N., Yagyu, T., Kinoshita, T., and Sasada, K. (2001). Source localization of EEG activity during hypnotically induced anxiety and relaxation. *International Journal of Psychophysiology*, 41(2), 143–153.

Jennings, N.R., Moreau, L., Nicholson, D., Ramchurn, S., Roberts, S., Rodden, T., and Rogers, A. (2014). Human–Agent Collectives. *Communications of the ACM*, 57(12), 80–88.

Jian, J.Y., Bisantz, A.M., and Drury, C.G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53–71.

Jonker, C.M. and Treur, J. (1999). Formal analysis of models for the dynamics of trust based on experiences. In *European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, 221–231. Springer Berlin Heidelberg.

Khawaji, A., Zhou, J., Chen, F., and Marcus, N. (2015). Using galvanic skin response (GSR) to measure trust and cognitive load in the text-chat environment. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 1989–1994. ACM.

Knyazev, G.G., Slobodskaya, H.R., and Wilson, G.D. (2002). Psychophysiological correlates of behavioural inhibition and activation. *Personality and Individual Differences*, 33(4), 647–660.

Lee, J. and Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243–1270.

Lee, J.D. and See, K.A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1), 50–80.

Long, Y., Jiang, X., and Zhou, X. (2012). To believe or not to believe: trust choice modulates brain responses in outcome evaluation. *Neuroscience*, 200, 50–58.

Lotte, F., Congedo, M., Lcuyer, A., Lamarche, F., and Arnaldi, B. (2007). A review of classification algorithms for EEG–based brain–computer interfaces. *Journal of Neural Engineering*, 4(2), R1–R13.

Ma, Q., Meng, L., and Shen, Q. (2015). You have my word: reciprocity expectation modulates feedback-related negativity in the trust game. *PloS one*, 10(2).

Mcfarland, D., Anderson, C., Muller, K.R., Schlogl, A., and Krusienski, D. (2006). BCI Meeting 2005-Workshop on BCI Signal Processing: Feature Extraction and Translation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2), 135–138.

Pfurtscheller, G., Flotzinger, D., and Kalcher, J. (1993). Brain-computer interface–a new communication device for handicapped persons. *Journal of Microcomputer Applications*, 16(3), 293 – 299.

Rakotomamonjy, A., Guigue, V., Mallet, G., and Alvarado, V. (2005). Ensemble of SVMs for improving brain computer interface p300 speller performances. In *International conference on artificial neural networks*, 45–50. Springer Berlin Heidelberg.

Ray, W. and Cole, H. (1985). EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes. *Science*, 228(4700), 750–752.

Riedl, R. and Javor, A. (2012). The biology of trust: Integrating evidence from genetics, endocrinology, and functional brain imaging. *Journal of Neuroscience, Psychology, and Economics*, 5(2), 63–91.

Righi, S., Mecacci, L., and Viggiano, M.P. (2009). Anxiety, cognitive self-evaluation and performance: ERP correlates. *Journal of Anxiety Disorders*, 23(8), 1132–1138.

Sadrfaridpour, B., Saeidi, H., Burke, J., Madathil, K., and Wang, Y. (2016). *Modeling and Control of Trust in Human-Robot Collaborative Manufacturing*, 115–141. Springer US, Boston, MA.

Sheridan, T.B. and Parasuraman, R. (2005). Human–automation interaction. *Reviews of human factors and ergonomics*, 1(1), 89–129.

Theodoridis, S. and Koutroumbas, K. (2006). *Pattern Recognition*. Pattern Recognition Series. Elsevier Science.

Yagoda, R.E. and Gillan, D.J. (2012). You Want Me to Trust a ROBOT? The Development of a Human-Robot Interaction Trust Scale. *International Journal of Social Robotics*, 4(3), 235–248.