2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 11-14, 2020. Toronto, Canada

Deceit Detection: Identification of Presenter's Subjective Doubt Using Affective Observation Neural Network Analysis

Xuanying Zhu
Research School of Computer Science
The Australian National University
Canberra, Australia
xuanying.zhu@anu.edu.au

Tom Gedeon
Research School of Computer Science
The Australian National University
Canberra, Australia
tom@cs.anu.edu.au

Sabrina Caldwell
Research School of Computer Science
The Australian National University
Canberra, Australia
sabrina.caldwell@anu.edu.au

Richard Jones
Research School of Computer Science
The Australian National University
Canberra, Australia
richard.jones@cs.anu.edu.au

Xiaohan Gu
Research School of Computer Science
The Australian National University
Canberra, Australia
xiaohan.gu@anu.edu.au

***Abstract*—We live in a world surrounded with 'fake news' and manipulated information, so a system assisting people with knowing what information to trust would be beneficial. Our research investigates situations where the *presenters* themselves have doubts about the information they are delivering, and we detect this via advanced affective computing techniques. To this end we examine the physiological foundations for observer recognition of the *doubt effect*: the subjective belief or disbelief of a presenter in some information he or she is presenting. Firstly, we construct stimulus videos that display presenters delivering information about which we manipulate their degree of doubt. We then show these stimuli to observers, and record four of their physiological signals. We find that a generalised neural network trained with physiological features is more accurate in differentiating the presenters' doubt/manipulated belief when compared with the same observers' own conscious judgments. The affective recognition performance improves when we analyse the physiological signals using multi-task learning techniques to train personalised and group personalised neural networks. The ability to recognise this *doubt effect* derives from observers' fundamental emotional reactions to the viewed stimuli, reflected in their physiological responses, and learnt by our neural networks. We believe this system using observer physiological signals collected in real life could reveal accurate and hidden audience distrust, which could in turn lead to enhanced truthfulness in future public-presented statements.**

Keywords*—*Neural networks (NN), Multitask learning (MTL), Blood volume pulse, Galvanic skin response, Skin temperature, Pupillary dilation, Information veracity, Doubt, Trust, Subjective belief, Presenters, Audiences

I. Introduction

People learn, cooperate, and socially bond through communication. To smooth the learning process, facilitate collaboration and maintain enduring social bonds, information and knowledge being shared and exchanged should be honest and faithful [1] so that people can navigate the information and knowledge communicated with confidence and trust. However, development of information technologies has revolutionised the way communications are produced and distributed, with the side-effect of facilitating the proliferation of manipulated information. The use of incorrect or exaggerated product statements for commercial benefit has been widespread in advertising for many decades [2]. In recent times, on social media, the increasing prevalence of false stories, such as fake news, is also concerning [3]. Credibility of information communication becomes problematic under the weight of manipulated information and it could potentially cause the people being deceived to suffer from devastating consequences in their personal lives [3]. Therefore, the skill of knowing whom and what to trust is essential to assist people in avoiding being deceived. [4].

However, people detect deception consciously at only around chance levels [5]. At least this is the case when people are asked to provide direct cognitive judgments about the dishonesty of others. Yet, when people's judgments are assessed indirectly through their non-cognitive signals, they do seem to be better at distinguishing liars from truth-tellers, even though they may not be consciously aware that they are being deceived [6]. DePaulo et al. [7] and Albrechtsen et al. [8] found that people's instant and intuitive judgements seem to distinguish liars from truth-tellers better than their slow and deliberative judgements made after conscious reasoning. This may be because some less apparent cues of deceit given away by a person with the intent to deceive may be unconsciously picked up by observers, alerting them to potential threats [1]. People being unconsciously affected by subtle deceiving cues may not

978-1-7281-8526-2/20/$31.00 ©2020 IEEE

experience obvious changes in their mental processes, but since human feelings are always accompanied by physiological changes [9], their physiological responses cause measurable reactions. Thus, physiological signals from people who are being deceived may act as a deceit detector.

Consistent aberrations of physiological response have been observed in liars. Since lying requires stressful cognitive processes to suppress the truth while making counterfactual statements, it is cognitively and emotionally taxing, resulting in increased sympathetic nervous system (SNS) activity [10]. As a result, liars' galvanic skin response, pupil dilation and heart rate increase [11], [12], and these effects could subtly affect behaviour, posture, or dynamics of movement and thus be visible to observers. People being deceived are sensitive to the subtle cognitive and emotional cues conveyed by liars and in response have similar physiological reactions [1]. For example, it has been found that when observing a liar, participants have lowered skin temperature on their fingers and greater pupillary responses [13]. Furthermore, in [14], people were found to have increased fixations and durations of eye gaze attending to manipulated areas of images. Higher heart rate was also found in people who watched deceptive events [15].

Despite physiological correlates of intentional deceit being examined previously, we could find no similar work to our research into inadvertent lying, a more subtle form of deceit. Our previous work explored the feasibility of using observers' pupil dilation to detect whether a person's subjective belief in some information has been manipulated [16]. The manipulated subjective belief refers to the situation where the presenter has cause to doubt the information they are presenting, but he or she is not explicitly intending to deceive, and so is conveying a subtle form of dishonest information. We found that neural networks trained with statistical features of observers' pupillary size reached a higher accuracy in differentiating the manipulated subjective belief when compared to each observer's own veracity judgments.

Here, we extend this work to ascertain whether neural networks trained on observers' other physiological signals such as blood volume pulse (BVP), galvanic skin responses (GSR), and skin temperature (ST) can identify presenters' subjective belief. Also, as it has been shown that a combination of multiple physiological signals is better at recognising an individual's affective state than a single physiological signal [9], we examine whether an integration of multiple physiological signals from observers is more accurate at detecting presenters' subjective belief or doubt. Additionally, since people may have different physiological responses to the same emotional stimulus [17], [18], we investigate personalisation of this subjective belief detection for individuals and groups by adopting a multi-task learning (MTL) approach. We seek to discover if the analysis of physiological indicators from observers and clusters of observers watching deceiving presenters would indicate the presence of subjective doubt on the part of the presenters. If so, this approach may reveal hidden audience distrust, help people notice their distrust, and ultimately could lead to increased truthfulness in public messaging.

II. Experimental design

This experiment uses observers' blood volume pulse (BVP), galvanic skin response (GSR), skin temperature (ST) and pupillary dilation (PD) signals to detect presenters' (manipulated) subjective belief in the information they present. To create experiment materials, following a similar experimental design to our pilot study [16], we first constructed eight extracts formatted as book publisher advertising material. We then recorded thirty-two videos, each of which contains a presenter reading out one of the constructed book extracts. Subsequently, we recruited some participants as observers to watch the recorded video stimuli while we recorded their BVP, GSR, ST and PD. We also collected observers' conscious judgment of the presenters' belief in the content presented in the videos via a survey. A schematic diagram of the experiment setup is shown in Fig. 1.

A. Book Extracts Construction

Eight extracts from books or web pages [19]–[25] were constructed, phrased and formatted like book publisher advertising materials. The contents of these extracts fall into four commonly discussed topics, namely health, astronomy, humanities and lifestyle, to ensure each participant has similar familiarity with the contents overall. For each topic, we constructed an extract based on the real content of a relevant book or webpage, or based on topics within the realm of common knowledge yet controversial enough to allow for the doubt effect. We then constructed another cognate extract by manipulating the content of another relevant book, webpage, or common knowledge. In all, we constructed a total of eight extracts which (to the authors) appear of similar plausibility in terms of their contents.

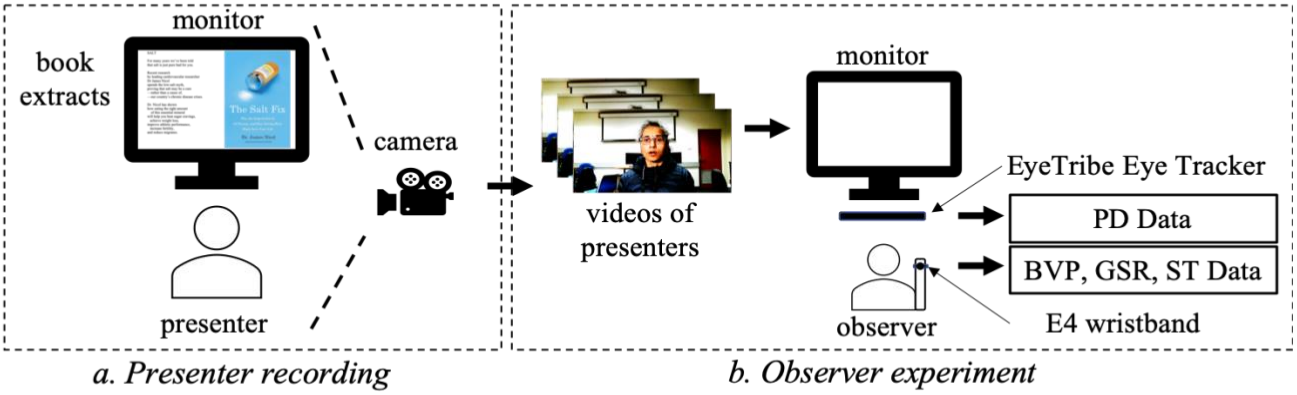

Fig. 1. A schematic diagram of our experiment setup

B. *Presenter Recording: Creating Video Stimuli*

Thirty-two videos were subsequently recorded, each of which involves an individual presenting one of the above-mentioned extracts. To record these video stimuli, eight presenters, four males and four females, were recruited as actors, with Ethics Approval obtained from the Australian National University Human Research Ethics Committee. Their task was to read out given extracts smoothly. None of these presenters were professional actors. We also ensure for each age group from 18-24, 25-39, 40-49, and 50+, there is one male and one female presenter to maintain age and gender balance. To minimize the influence of topic familiarity for presenters, each presenter only read one extract from each topic and thus presented four extracts covering four topics in total.

After giving their informed consent, presenters first presented two book extracts (selected in an order balanced fashion) which they could presume to be true (naive subjective belief condition). In the brief period before presenting the third and fourth extracts, they were told "Sorry the next two are a bit bogus. Please present them anyway". They then presented the other two extracts (manipulated subjective belief condition). While they were presenting the four extracts, a camera was placed at 1 meter from the actors and filmed them from the chest up. In this way, we recorded a total of sixteen 35–60 sec videos in which presenters' subjective belief in the topics have been manipulated to induce subjective disbelief, and sixteen videos where their subjective belief is unchanged. These videos were subsequently shown to observers.

C. *Observer Experiment: Judging manipulated belief*

1) Participants and Procedure: Thirty participants were recruited as observers, with Ethics Approval obtained from the Australian National University Human Research Ethics Committee. Seven participants were excluded due to one or more predefined exclusion criteria: being acquainted with the video presenters, having a history of cardiovascular disease, and technical failures of sensors. The final observer samples consist of twenty-three participants, thirteen males and ten females, from 18 to 24 years in age (average = 21.2, standard deviation = 2.0).

The observer experiment was conducted with each individual participant in the same quiet experiment room. Participants were forewarned that their goal during the experiment was to identify the veracity of the presented content. This is because dishonest information has a higher chance of getting caught when observers are alerted to the possibility of dishonesty in advance [13], which may also apply to manipulated (subjective doubt) beliefs. After providing written informed consent, an Empatica E4 watch [26] was attached to participants on the wrist of their non-dominant hand to collect physiological signals. An EyeTribe eye tracker [27] was also placed in front of them to collect their pupillary responses. After filling in a questionnaire to collect demographic and health information that may affect physiological responses, participants then watched sixteen presenter videos one by one and were asked to provide responses to questions of "Do you feel the presenter is a trustworthy person?" and "Do you think the presenter trusts what he/she said?" on a binary scale (yes, or no). The videos were presented in an order balanced way to avoid effects of presentation sequencing.

2) Sensor Recording: We collected four physiological signals from observers, namely BVP, GSR, ST and PD as well as a derived physiological signal, HR.

a) Blood Volume Pulse (BVP): indicates the volume of blood running through the vessels over a given period, affected by SNS activation in response to emotional reactions [28]. For example, BVP is found to be positively correlated with sadness and negatively correlated with stress [29]. Other cardiovascular measures, such as heart rate (HR) and heart rate variability (HRV) can be derived from BVP, and these are useful predictors of emotional valence [30]. In this experiment, we used an Empatica E4 wristband to collect BVP with a sampling rate of 64Hz [26].

b) Galvanic Skin Response (GSR): measures the amount of sweat on a person' skin. It contains a slow-moving tonic component, showing general activity of perspiratory glands caused by body or external temperature, and a faster-responding phasic component which is linearly correlated with the intensity of arousal in emotional states [31]. In this study, we recorded observers' GSR using an Empatica E4 wristband with a sampling rate of 4Hz.

c) Skin Temperature (ST): fluctuates due to vasodilatation of blood vessels induced by increased SNS activity [32]. It is found to be negatively correlated with threat emotions such as stress [33] and fear [34]. In this study, we recorded observers' wrist ST using an Empatica E4 wristband with a sampling rate of 4Hz

d) Pupillary Dilation (PD): provides an indication of changes and strengths in mental activites, and has been found to be correlated with emotionally engaging stimuli [35]. In this study, we used an EyeTribe eye tracker to capture pupil size at a sampling rate of 60Hz.

After data collection, 193 manipulated and 175 unmanipulated subjective belief observations were obtained. In the rest of this paper we will mostly use *doubting* and *trusting* to succinctly express presenters' manipulated subjective belief and unmanipulated subjective belief, respectively.

III. Methodology

To analyse observers' physiological responses, we first pre-processed each physiological signal with segmentation, noise removal and normalisation before we extracted features. We then trained an MTL neural network (NN) to customise a model for each type of person to assess presenters' doubting and trusting subjective belief.

A. *Pre-processing*

Transient noise was observed in the raw physiological signals due to participant movement, which mostly happened at the beginning and end of the recording when they filled in the demographic questionnaire and post-experiment survey. For all participants and all signals, we first extracted the subset of raw signal data recorded as participants watched presenter videos. Cubic spline interpolation was then applied to construct missing pupil size data caused by occasional eye blinks [36]. This procedure was employed on the pupil data of left and right eyes separately.

Physiological signals are individual-dependent. This means that different individuals may have the same physiological signal in different ranges. To reduce between-participant differences, we applied a min-max scaler to all physiological signals separately, scaling signals to a range between 0 and 1.

After normalisation, we smoothed the signals to remove noise artefacts. For BVP, GSR and ST, we used a lowpass Butterworth filter with an order of 6 and a cut-off frequency of 0.5 Hz, 0.2 Hz [37] and 0.3 Hz [38] respectively to form a low-passed (LP) BVP, GSR and ST signal. For PD, we applied a 10-point Hann moving window average to left pupil and right pupil data separately [36].

Following this, we segmented both the normalised signals and filtered signals by each video watching session, so that each segmented physiological data set corresponds to one observer's physiological state evoked by the experience of watching one video.

B. Features Extraction

After pre-processing the raw signals, we generated time- and frequency-domain features that characterise the changes in the physiological signals over the time participants spent on watching each video.

1) BVP features: Following [39] which uses BVP for emotion recognition, we first calculated the following eight time-domain features from the normalised and LP BVP.

- *Minimum*
- *Maximum*
- *Mean*
- *Standard Deviation*
- *Variance*
- *Root Mean Square*
- *Means of the Absolute Values of the First Difference*
- *Means of the Absolute Values of the Second Difference*

Following [40], we derived heart rate (HR) from BVP by identifying systolic peaks from BVP. We then calculated the above-mentioned eight time-domain features from HR. Additionally, we extracted another eight time-domain features from HR which were shown to be correlated with external stimuli [41].

- *Inter Beats Interval (IBI)*
- *Average Beats per Minute (BPM)*
- *Standard Deviation of Intervals between Heart Beats (SDNN)*
- *Standard Deviation of Differences between Adjacent R-R Intervals (SDSD)*
- *Root Mean Square of Differences between Adjacent R-R Intervals (rMSSD)*
- *Percentage of differences greater than 20ms (pNN20)*
- *Percentage of differences greater than 50ms (pNN50)*
- *Proportion of differences greater than 50ms / 20ms (pNN50/pNN20)*

2) GSR features: For normalised and LP GSR, we first calculated sixteen time-domain features. For each of the normalised and LP GSR, we calculated minimum, maximum, mean, standard deviation, variance, root mean square, and means of the absolute values of the first and second differences.

To extract the DC component of GSR [31], we applied a very low pass Butterworth filter with a cut-off frequency of 0.08 Hz to derive the Very Low Pass signal (VLP). We additionally obtained a detrended SCR signal without DC component by removing continuous piecewise linear trend in both LP and VLP GSR. Subsequently, we calculated the following seven features:

- *Number of SCR occurences for VLP, LP and normalised GSR (3 features)*
- *Mean of amplitudes of SCRs in VLP, LP and normalised GSR (3 features)*
- *Ratio of SCR occurrences in VLP to occurrences in LP*

3) ST features: Similarly to the GSR signal, we first calculated eight time-domain features, including minimum, maximum, mean, standard deviation, variance, root mean square, and means of the absolute values of the first and second differences for the normalised and LP ST. Then we used a very low pass Butterworth filter with a cut-off frequency of 0.08 Hz to the normalised ST to form the VLP ST signal. We computed the numbers and the mean of amplitudes of peaks for VLP and LP ST signals as well as the ratio of peaks in VLP to those in LP as features.

4) PD features: For normalised left, right pupillary size, and the averaged pupillary size of left and right eyes, the minimum, maximum, mean, standard deviation, variance, root mean square, and means of the absolute values of the first and second difference were calculated as features. We then used a very low pass Butterworth filter with a cut-off frequency of 0.08 Hz to the normalised left, right and averaged PD signal to form the left VLP PD and right VLP PD signal. The same filter was also applied to the averaged pupillary size of left and right eye to form the average VLP PD signal. We then extracted numbers and amplitudes of peak occurences for left, right and average VLP and LP PD signals as well as the ratio of peak occurances in VLP to those in LP for the left, right and average signals.

From these data we collected a total of 119 features across the four physiological signals: 34 (BVP) + 23 (GSR) + 23 (ST) + 39 (PD).

C. Features Selection

While large numbers of features can be derived from physiological signals to make predictions, this full set of features may include irrelevant features. These redundant features may outweigh the more effective features and affect classification performance, especially for a small dataset. Hence, we applied feature selection to reduce the chance of overfitting, along with early stopping. Inspired by [42] where classifiers trained with subsets of physiological features selected by Genetic Algorithms (GAs) outperform other feature selection methods, in this work we used a GA for feature selection.

We set the initial population for the GA to use all features, and we set a chromosome as a binary string where the index for each bit represents a specific feature, and the value indicates if the feature is used for classification. The optimisation goal of the GA is to find better subsets of features as candidate chromosomes by determining the presence or absence of every possible feature in the model, using the performance of a classifier as the fitness function.

D. Classification

In this study, we were interested in determining the trusting and doubting subjective beliefs of presenters using a combination of observer BVP, GSR, ST, and PD measurements as monitored signals. Classification was attempted via NN, a nonlinear classifier containing several hidden layers, each of which performs a non-linear transformation $x_{i+1} = \sigma(w_i x_i + b_i)$ where x_i is the input of the i th layer, w_i and $b_i$are the weight matrix and bias, and σ is the activation function.

To evaluate the difference between the generalized veracity detection method that does not take individual participant differences into account and personalized veracity detection, we performed classification in three different ways by building 1) generalised, 2) personalised, and 3) group-personalised veracity detection models.

1) Generalised Veracity Model: In the first approach, a generalized NN was trained using a leave-one-participant-out validation scheme, in which data from one participant were used as testing data and data for remaining participants were treated as training data. The NN was a fully connected neural network with a sigmoid hidden layer of size 512 and an output layer of two output neurons, representing the trusting and doubting subjective beliefs of presenters. The number of hidden neurons was set to 512 after we tested our neural networks with different hidden neuron size from 10 to 1024 and found 512 to be optimal. The NN was trained with the Adam optimizer [43] using backpropagation with the Cross-Entropy loss function.

2) Personalised Veracity Model: Since people have been found to have different reactions to the same stimulus [17], [18], a generic model trained to estimate presenters' subjective belief is limited in performance. Therefore in the second approach, we trained a multi-task learning neural network (MTL-NN) to account for inter-individual variability. A MTL-NN solves multiple tasks simultaneously with a shared representation of the tasks. In other words, a MTL-NN contains several initial layers shared across all tasks, and N task-specific classifiers, one for each task, where N is the number of tasks. The optimization of loss functions is done concurrently by switching between different tasks.

As depicted in Fig. 2, in this model, we treated assessing the subjective belief of a viewed presenter for an observer as a single task. For shared layers, we used two fully-connected sigmoid layers with 350 neurons. The participant-specific classifiers contain a fully-connected sigmoid layer with 50 hidden neurons and an output layer with 2 output neurons, which indicate the trusting or doubting subjective beliefs of presenters. The Cross-Entropy was optimised as an objective function with Adam optimiser [43]. For each participant, a random 80%-20% split was used to partition data into training and testing sets.

3) Group Personalised Veracity Model: The second approach above is only valid when each person has sufficient labelled data. Moreover, one major limitation of such a method is that it cannot generalize to new users. Thus, inspired from [44] where MTL was used to predict people's mood by their personality and gender, and from [17], [18] in which people with different gender and age were found to be affected differently by the same stimulus, we first clustered observers by their gender, age and ethinicity, and treated estimating viewed presenters' subjective belief for a given cluster as one task. This method can deal with new users by assigning them to appropriate clusters based on their gender and age. We applied a K-means clustering algorithm to observers' gender and age, and assess the quality of clustering using silhouette score [45], a measure of how similar each sample is to its own cluster compared to other clusters. In this study, the highest silhouette score was achieved when K = 3.

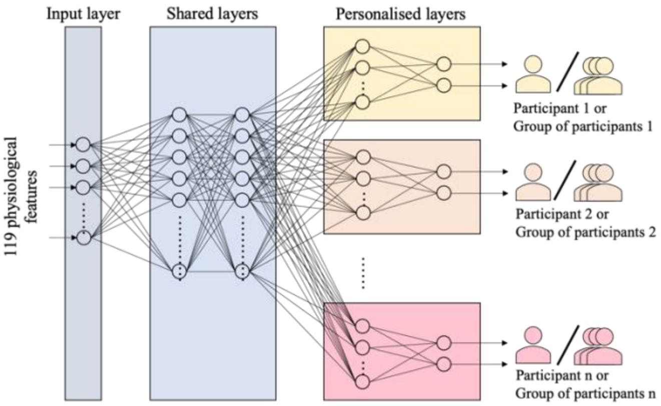

Fig. 2. Multi-task learning neural network architecture with two shared layers, and a participant- / group-specific layer.

Then we built a MTL-NN with K participant-specific classifiers, one for each group of participants, as Fig. 2 shows. We treated assessing the subjective belief of presenters for a group of similar observers based on gender and age as one single task. We used two fully-connected sigmoid layers with 350 neurons for shared layers and one full-connected sigmoid hidden layer with 100 hidden neurons plus a fully-connected output layer with 2 output neurons for each group-specific layer. The model was trained with the Adam optimiser [43] and cross-entropy was used as an objective function. For each participant, a random 80%-20% split was used to partition data into training and testing sets.

IV. Results and discussion

To evaluate the effectiveness of our models, we used *accuracy*, *precision*, *recall*, and *F1-score* as evaluation measures. Performance of all three models was calculated based on the average results of 10 runs.

A. Observers' Veracity Judgments

We first examined how well our observers were able to distinguish between the trusting and doubting subjective belief of presenters in videos. As listed in TABLE I, the overall accuracy was 52%, slightly above the chance level of 50%. This is consistent with the literature [5] and with earlier findings of the accuracy of people's conscious judgments of the veracity of smiles [36], anger [46], and of depression levels [47], which are all only marginally higher than chance levels. This lack of recognition accuracy could even have been exacerbated if our presenter participants were recruited from an actor population who had previously received sufficient training in respect of promoting products with exaggerated or misleading statements. Future research should explore the accuracy of conscious judgments from observers who viewed videos of experienced actors as presenters.

The average ratios of consciously distinguishing doubting and trusting subjective belief correctly were 55% and 48%.

B. Generalised Classification

We first trained the generalised NN using a leave-one-participant-out validation scheme. We also tested two standard classifiers as a baseline: Random Forest (RF) and support vector machine (SVM) with linear kernel. In addition, since NNs benefit from large input size, we tested the performance of these models on the full set of 119 features. The veracity detection performance of these models is summarised in TABLE II, and the performance of our generalised model on different beliefs can be found in TABLE III.

As shown in TABLE II, our generalised veracity model achieved a classification accuracy of 63% with all features, while RF and SVM reached 59% and 58% with a subset of features selected by GA, respectively. Statistical analysis was conducted on the results using the Student's t-test since different models trained with physiological features share normality and equality of variance across comparison groups. In accordance with the Student's t-test, our generalised veracity model performed significantly better than RF and SVM ($p < 0.005$). Also, similar to [44], our generalised NN performed better with a full set of features than with features selected by GA ($p<0.01$). This may imply that NN compared to RF and SVM may be a more effective model to detect observer responses to the veracity of others. The most effective feature set is the full set of features. Therefore, the remaining two models, personalised classifier and group personalised classifier, are trained on NNs with all features.

The result achieved by our generalised model was also statistically significantly better when compared to the null hypothesis that each class has equal chance of being selected by the classifier ($p < 0.01$) and when compared to observers' conscious judgments shown in TABLE I ($p < 0.05$). It could imply that although humans are not good at consciously detecting the subjective belief of others, in general they can emotionally sense deception in the content. Observers' physiological changes due to presenters' deceiving behaviours can be detected by computational classifiers (such as neural networks). In other studies which examined the veracity of two basic emotions [36], [46], it has also been found that human unconscious physiological response is better than their conscious judgment. Taken together, it could suggest that unconscious responses from instinctive human ability, which has been adaptively evolved by natural selection, can make use of cues to identify deceiving individuals without being influenced by conscious biases.

TABLE I. RESULTS OF OBSERVERS SUBJECTIVE VERACITY JUDGMENTS ON PRESENTERS BELIEF

Presenter Belief	Observer Subjective Judgment		
	Precision	*Recall*	*F1 score*
Doubting	0.52	0.55	0.53
Trusting	0.52	0.48	0.5
Average	0.52	0.52	0.52
Overall Accuracy	0.52		

TABLE II. PERFORMANCES OF GENERALISED VERACITY DETECTION MODEL

Classifier	*Accuracy*	*Precision*	*Recall*	*F1 score*
RF	0.59	0.59	0.58	0.59
RF (all features)	0.56	0.56	0.55	0.55
SVM	0.58	0.58	0.58	0.58
SVM (all features)	0.55	0.54	0.55	0.55
Our generalised NN	0.61	0.61	0.61	0.60
Our generalised NN (all features)	0.63	0.64	0.64	0.63

However, as TABLE III shows, in the doubting condition our generalised model only outperformed observers' random subjective judgements by 3% in recall ($p<0.005$), meaning that 42% of doubting observations could not be correctly estimated by our generalised model. Similarly, under the trusting condition, the generalised model could only identify 59% of the trusting observations accurately. This could be due to varying patterns of physiological responses from observers evoked by the same stimulus, and thus a generalised model trained on a population-based approach may not be optimal.

C. Personalised Classification

Taking individual differences into account, the second classification approach trained an MTL-NN where 23 participant-specific classifiers were built, one for each observer. For each participant, a random 80%-20% split was used to partition data into training and testing sets. The overall classification results are listed in TABLE IV.

Personalised classification resulted in an overall accuracy of 68% and an average precision, recall and F1 score of 72%, 74% and 72% respectively. All results were statistically significant compared to chance level classification for all observers ($p<0.01$), and generalised classification ($p<0.01$). For both the doubting and trusting conditions, personalised classification outperforms the generalised model in all measures ($p<0.01$ in all cases). The personalised model can recognise more doubting and trusting observations correctly than the generalised model; an increase of at least 12% was obtained on recall in the doubting condition and on precision in the trusting

TABLE III. OVERALL PERFORMANCES OF GENERALISED VERACITY DETECTION MODEL ON DOUBTING AND TRUSTING CONDITION

Presenter Belief	Generalised Veracity Model		
	Precision	*Recall*	*F1 score*
Doubting	0.68	0.58	0.62
Trusting	0.59	0.70	0.64
Average	0.64	0.64	0.63
Overall Accuracy	0.63		

TABLE IV. OVERALL PERFORMANCES OF PERSONALISED VERACITY DETECTION MODEL.

Presenter Belief	Personalised Veracity Model		
	Precision	*Recall*	*F1 score*
Doubting	0.71	0.74	0.73
Trusting	0.70	0.74	0.7
Average	0.72	0.74	0.72
Overall Accuracy	0.68		

condition. Therefore, the personalised model is more effective in estimating subjective belief of other individuals than detectors trained on population averages.

However, on an individual basis, it has been observed that the effectiveness of the personalised classifier varied among different observers. One possible explanation is that since observers with different age, gender and ethnicity seem to respond differently towards the same stimulus, they may have varying patterns of physiological signals corresponding to the same manipulated information [17], [18]. Also, this could be because the number of viewed sessions for each observer in this study may not be sufficient to build a robust model on an individual scale. Future study could examine the minimal number of presenters' videos viewed by each observer required to obtain a highly accurate observer-based classifier.

D. *Group Personalised Classification*

To validate whether observers' attributes create a difference in their physiological responses to manipulated subjective belief, we trained a group personalised veracity classifier by first clustering observers by their gender, age and ethnicity and then building an MTL-NN with several group-specific classifiers, one for each group. For each group of participants, a random 80%-20% split was used to partition data into training and testing sets. The overall classification results are provided in TABLE V, and performances of the models on each group of observers are listed in TABLE VI.

This group-personalised model resulted in a mean accuracy of 88%. The average precision, recall and F1 score varied between 88%-89%. This is a very substantial improvement over the close-to-chance results from conscious choices, especially noting that physiological signals are highly noisy data.

When compared to personalised and generalised models, the group-personalised model is more accurate, indicated by higher overall accuracy and higher other measures in both doubting and trusting conditions ($p<0.01$ in all measure comparisons). As clearly seen from TABLE III, IV and V, using MTL to personalise NN models by multitasking over clusters of similar observers provides dramatic improvements to subjective belief veracity estimation performance. The improvement in accuracy over the non-personalised and personalised models is at least 20%. The increase in F1 score also indicates that the group-personalised approach can better recognise both trusting and doubting observations.

Despite the impressive performance of the group-personalised model, as listed in TABLE VI, there were differences in effectiveness on estimating subjective beliefs of presenters, with the second observer group (which forms a more diffuse cluster than the other two) being lower than the other two groups, though still well above chance. This could imply that besides observers' age, gender and ethnicity that have been considered, there might be other factors impacting the capability of veracity models trained on physiological responses from groups of observers. For example, as emotional responses to stimuli can depend on personality [48] and familiarity towards stimuli [49], future work could examine the impact of observers' personality on the group-personalised veracity model.

TABLE V. OVERALL PERFORMANCES FOR GROUP PERSONALISED VERACITY DETECTION MODEL.

Presenter Belief	**Group Personalised Veracity Model**		
	Precision	***Recall***	***F1 score***
Doubting	0.88	0.91	0.89
Trusting	0.91	0.88	0.89
Average	0.89	0.88	0.89
Overall Accuracy	0.88		

TABLE VI. PERFORMANCES OF GROUP PERSONALISED VERACITY DETECTION MODEL ON EACH GROUP OF OBSERVERS

Group	**Group Personalised Veracity Model**			
	Accuracy	***Precision***	***Recall***	***F1 score***
1	0.91	0.91	0.91	0.9
2	0.84	0.85	0.85	0.84
3	0.90	0.92	0.92	0.92
Average	0.88	0.89	0.89	0.89

V. CONCLUSIONS AND FUTURE WORK

Our work explored physiological signals from observers to detect the doubt effect where a presenter's subjective belief in some content manipulated. We showed that a generalised NN trained on a population base reached a higher accuracy in differentiating doubting and trusting information compared with the conscious veracity judgments from the same observers. This recognition was significantly improved when MTL was used to account for individual differences or group differences. This is attributable to the ability of MTL which can both allow each individual to have a model customised for them and share data of other people through shared hidden layers.

Presenters recruited in this study are naïve individuals with no acting experience. In future work, some actors could be recruited as presenters to examine the effect of acting expertise to the ability for observers to differentiate true statement from fake or exaggerated information. Attributes of observers, such as their personality, could be collected to investigate further group distinctions for group differences classifications. Additional activity such as gestures [50] could also be tracked. With increasing amounts of data collected from wider groups of observers, stronger conclusions may be drawn in subsequent studies and allow the use of recent deep learning models, such as Convolutional Neural Networks or Long Short-Term Memory, which may achieve more accurate recognition results.

REFERENCES

[1] L. ten Brinke, K. D. Vohs, and D. R. Carney, "Can ordinary people detect deception after all?," Trends Cogn. Sci., vol. 20, no. 8, pp. 579–588, 2016.

[2] A. Rhodes and C. M. Wilson, "False advertising," RAND J. Econ., vol. 49, no. 2, pp. 348–369, 2018.

[3] M. Tsikerdekis and S. Zeadally, "Online deception in social media," Commun. ACM, vol. 57, no. 9, pp. 72–80, 2014.

[4] W. Von Hippel and R. Trivers, "The evolution and psychology of self-deception," Behav. Brain Sci., vol. 34, no. 1, pp. 1–16, 2011.

[5] C. F. Bond Jr and B. M. DePaulo, "Accuracy of deception judgments," Personal. Soc. Psychol. Rev., vol. 10, no. 3, pp. 214–234, 2006.

[6] J. Ulatowska, "Different questions--different accuracy? The accuracy of various indirect question types in deception detection," Psychiatry, Psychol. Law, vol. 21, no. 2, pp. 231–240, 2014.

[7] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception.," Psychol. Bull., vol. 129,

no. 1, p. 74, 2003.

[8] J. S. Albrechtsen, C. A. Meissner, and K. J. Susa, “Can intuition improve deception detection performance?,” J. Exp. Soc. Psychol., vol. 45, no. 4, pp. 1052–1055, 2009.

[9] W. Ambach and M. Gamer, “Physiological Measures in the Detection of Deception and Concealed Information,” Detect. Concealed Inf. Decept., p. 1, 2018.

[10] K.-H. Jung and J.-H. Lee, “Cognitive and emotional correlates of different types of deception,” Soc. Behav. Personal. an Int. J., vol. 40, no. 4, pp. 575–584, 2012.

[11] S. Ströfer, M. L. Noordzij, E. G. Ufkes, and E. Giebels, “Deceptive intentions: Can cues to deception be measured before a lie is even stated?,” PLoS One, vol. 10, no. 5, p. e0125237, 2015.

[12] A. Elkins, S. Zafeiriou, M. Pantic, and J. Burgoon, “Unobtrusive deception detection,” in The Oxford handbook of affective computing, Oxford Univ. Press, 2014.

[13] A. van’t Veer, “Effortless morality: Cognitive and affective processes in deception and its detection,” Diss. Tilbg. Univ., 2016.

[14] S. Caldwell, T. Gedeon, R. Jones, and L. Copeland, “Imperfect Understandings: A Grounded Theory And Eye Gaze Investigation Of Human Perceptions Of Manipulated And Unmanipulated Digital Images,” in Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science, 2015, vol. 308.

[15] G. Duran, I. Tapiero, and G. A. Michael, “Resting heart rate: A physiological predicator of lie detection ability,” Physiol. Behav., vol. 186, pp. 10–15, 2018.

[16] X. Zhu, Z. Qin, T. Gedeon, R. Jones, M. Z. Hossain, and S. Caldwell, “Detecting the Doubt Effect and Subjective Beliefs Using Neural Networks and Observers’ Pupillary Responses,” in International Conference on Neural Information Processing, 2018, pp. 610–621.

[17] M. Bianchin and A. Angrilli, “Gender differences in emotional responses: A psychophysiological study,” Physiol. Behav., vol. 105, no. 4, pp. 925–932, 2012.

[18] V. Orgeta, “Specificity of age differences in emotion regulation,” Aging Ment. Heal., vol. 13, no. 6, pp. 818–826, 2009.

[19] J. DiNicolantonio, The Salt Fix. Harmony, 2017.

[20] S. M. Stanford et al., “Diabetes reversal by inhibition of the low-molecular-weight tyrosine phosphatase,” Nat. Chem. Biol., vol. 13, no. 6, p. 624, 2017.

[21] C. Cassella, “We Just Got More Compelling Evidence That The Moon Is Loaded With Water,” Science Alert, 2018. .

[22] G. Lisa, “Curiosity finds that Mars’ methane changes with the seasons,” ScienceNews, 2018. [Online]. Available: https://www.sciencenews.org/article/curiosity-finds-mars-methane-changes-seasons.

[23] Y. Paramhansa, “Autobiography of a Yogi,” Ananda India, 2018. .

[24] D. N. Murphy, The Marlowe-Shakespeare Continuum. 2013.

[25] K. Miller, “Should you avoid bananas if you’re trying to lose weight?,” health24, 2017. .

[26] Empatica, “E4 wristband.” [Online]. Available: https://www.empatica.com/research/e4/. [Accessed: 30-May-2018].

[27] TheEyeTribe, “The EyeTribe.” [Online]. Available: http://theeyetribe.com/theeyetribe.com/about/index.html.

[28] K. Gouizi, F. Bereksi Reguig, and C. Maaoui, “Emotion recognition from physiological signals,” J. Med. Eng. Technol., vol. 35, no. 7, pp. 300–307, 2011.

[29] N. Sharma and T. Gedeon, “Modeling a stress signal,” Appl. Soft Comput., vol. 14, pp. 53–61, 2014.

[30] P. Ekman, “An argument for basic emotions,” Cogn. Emot., vol. 6, no. March 2014, pp. 37–41, 2008.

[31] J. Kim and E. André, “Emotion recognition based on physiological changes in music listening,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 12, pp. 2067–2083, 2008.

[32] R. M. Stern, W. J. Ray, and K. S. Quigley, Psychophysiological recording. Oxford University Press, USA, 2001.

[33] F. Al-Shargie, T. B. Tang, and M. Kiguchi, “Mental stress grading based on fNIRS signals,” in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2016, vol. 2016-Octob, pp. 5140–5143.

[34] J. E. LeDoux and S. G. Hofmann, “The subjective experience of emotion: a fearful view,” Curr. Opin. Behav. Sci., vol. 19, pp. 67–72, Feb. 2018.

[35] B. Laeng, S. Sirois, and G. Gredebäck, “Pupillometry: a window to the preconscious?,” Perspect. Psychol. Sci., vol. 7, no. 1, pp. 18–27, 2012.

[36] M. Z. Hossain and T. Gedeon, “Effect of Parameter Tuning at Distinguishing Between Real and Posed Smiles from Observers’ Physiological Features,” in International Conference on Neural Information Processing, 2017, pp. 839–850.

[37] A. Haag, S. Goronzy, P. Schaich, and J. Williams, “Emotion recognition using bio-sensors: First steps towards an automatic system,” in Tutorial and research workshop on affective dialogue systems, 2004, pp. 36–48.

[38] P. C. Schmid, M. S. Mast, D. Bombari, F. W. Mast, and J. S. Lobmaier, “How mood states affect information processing during facial emotion recognition: an eye tracking study,” Swiss J. Psychol., 2011.

[39] A. Kushki, J. Fairley, S. Merja, G. King, and T. Chau, “Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites,” Physiol. Meas., vol. 32, no. 10, p. 1529, 2011.

[40] X. Zhu, T. Gedeon, S. Caldwell, and R. Jones, “Visceral versus Verbal: Can We See Depression?,” Acta Polytech. Hungarica, vol. 16, no. 9, 2019.

[41] P. Van Gent, H. Farah, N. van Nes, and B. van Arem, “Analysing Noisy Driver Physiology Real-Time Using Off-the-Self Sensors: Heart Rate Analysis Software from the Taking the Fast Lane Project.,” J. Open Res. Softw., 2018.

[42] J. S. Rahman, T. Gedeon, S. Caldwell, R. Jones, M. Z. Hossain, and X. Zhu, “Melodious Micro-frissons: Detecting Music Genres from Skin Response,” in Proceedings of the International Joint Conference on Neural Networks, 2019, vol. 2019-July, pp. 1–8.

[43] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv Prepr. arXiv1412.6980, 2014.

[44] S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, “Personalized Multitask Learning for Predicting Tomorrow’s Mood, Stress, and Health,” IEEE Trans. Affect. Comput., 2017.

[45] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., vol. 20, pp. 53–65, 1987.

[46] L. Chen, T. Gedeon, M. Z. Hossain, and S. Caldwell, “Are you really angry?: detecting emotion veracity as a proposed tool for interaction,” in Proceedings of the 29th Australian Conference on Computer-Human Interaction, 2017, pp. 412–416.

[47] X. Zhu, T. Gedeon, S. Caldwell, and R. Jones, “Detecting emotional reactions to videos of depression,” in IEEE International Conference on Intelligent Engineering Systems, 2019.

[48] S. Zhao, G. Ding, J. Han, and Y. Gao, “Personality-Aware Personalized Emotion Recognition from Physiological Signals.,” in IJCAI, 2018, pp. 1660–1667.

[49] A. Kawakami, K. Furukawa, K. Katahira, K. Kamiyama, and K. Okanoya, “Relations between musical structures and perceived and felt emotions,” Music Percept. An Interdiscip. J., vol. 30, no. 4, pp. 407–417, 2013.

[50] R. Gravina and Q. Li, “Emotion-relevant activity recognition based on smart cushion using multi-sensor fusion,” Inf. Fusion, vol. 48, pp. 1–10, 2019.