Detecting emotional reactions to videos of depression

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

***Abstract*— We can physically sense depression in others, and this recognition can be detected by using neural networks to analyse our physiological responses to observing individuals with depression. The behaviour of 16 individuals suffering from various levels depression were shown in short videos to 12 experiment participants (observers) whose physiological signals we recorded. Consciously, depression is not normally interesting, so does not provoke strong conscious recognition, and hence barely over chance, at 27%. However, at emotional levels, depression is interesting, so provokes physiological reactions we can measure, leading to neural network classification of 92%.**

Keywords*— *Depression Detection, Physiological Signals, Galvanic Skin Response, Skin Temperature, Pupil Dilation

I. INTRODUCTION

Depression is an internalising mental disorder [1]. Different from usual mood fluctuations and short-lived emotional responses to daily events, depression is a serious chronic health condition. It is often accompanied by persistent feelings of sadness, loss of interest and enjoyment, low self-esteem, inability to cope with everyday responsibilities [1], and at its worst, suicidal attempts and behaviours [2]. Therefore, effective diagnosis and treatment of depression is key to prevent suicides [3], to improve the quality of life and mental health of affected individuals, families and communities, and to reduce socio-economic costs [1].

Due to the large variation in clinical characteristics of depressed patients and thus the lack of a laboratory test for depression [4], current diagnosis is subjective and time-consuming. Commonly used diagnosis methods are generally based on self-reported questionnaires such as the Beck Depression Index (BDI) [5] or clinician assisted interview style assessments such as the Hamilton Rating Scale for Depression (HAMD) [6], which score patients' depression level by the severity of their symptoms. These tools are biased in nature, as they heavily rely on patients' ability and sincerity to honestly reflect their symptoms, and on individual clinicians' experience and opinions. Since, by definition, depressed patients have a weakened attitude towards life, making them not willing to express (or even be aware of) their emotional symptoms [7], these diagnostic methods are not always reliable, especially when health-care providers have not received sufficient clinical training and practice [4]. Hence, more objective diagnostic aids are needed.

Recent advances made in affective computing technology demonstrate the possibility of using physiological signals to assist diagnosing depression [8], since physiological responses can indicate emotions without conscious awareness [9]. For example, depressed patients are found to have different eye gaze behaviours [10], lower Galvanic Skin Response (GSR) [11] and reduced Heart Rate Variability (HRV) [12]. These physiological features provide more uniform and quantitative criteria, and combined with machine learning technologies, can play an important role in providing an objective assessment for depression.

Our goal here is to investigate the possibility of using physiological signals from observers to identify others' depression level. As depression affects certain areas of the brain which results in universal observable behavioural patterns such as differences in facial expression, eye movements and gestures [10], these subtle cues could be noticed by observers, which is then reflected in observers' physiological signals. Our previous work has demonstrated the feasibility of using observers' physiological signals as an indicator of others' stress [13], happiness [14], anger [15] and deceiving behaviours [16] using neural networks. We hope the identification of universal physiological indicators from observers watching depressed individuals would assist with more objective and earlier diagnosis, which combined with the use of known effective treatments would decrease the burden for individuals and society.

We use Neural Networks (NNs) to recognise depression levels based on features derived from observers' physiological responses to others' depression. We use genetic algorithm (GA) to select subsets of features for optimizing depression classifications since GA has been successfully applied to select features from physiological signals [13].

This paper examines whether observers are responding physiologically to depressed individuals, and whether a classifier could be developed to recognise other individuals' depression level using observers' physiological signals. It details an experiment conducted to collect multiple physiological response signals from experiment participants who watched videos of people with various levels of depression. Approaches for depression recognition of video watchers are developed and discussed, including a method for selecting optimally useful features from the response signals. The paper concludes with a summary of the findings and suggests directions for future work.

II. EXPERIMENTAL DESIGN

A. Stimuli

We used videos from the 2014 Audio-Visual Emotion Challenge (AVEC 2014) dataset [17]. AVEC 2014 consists of

978-1-7281-1213-8/19/$31.00 ©2019 IEEE

300 webcam video recordings of participants individually completing either the task of reading aloud a paragraph (the Northwind task) or answering a set of questions (the Freeform task); both in German. Each recording was labelled with dimensions of valence, arousal and dominance, and a single depression level. The three affective dimensions were annotated continuously by a team of five naïve assessors for each video frame, and the actual depression level was derived from involved participants self-reported depression level indicated by the Beck Depression Inventory – II (BDI-II) [5]. This index gives depression scores ranging from 0 to 63 and groups the scores into four depression categories:

- 0-13: indicates no or minimal depression
- 14-19: indicates mild depression
- 20-28: indicates moderate depression
- 29-63: indicates severe depression

AVEC 2014 addresses the prediction of these affective dimensions and depression levels as two separate sub-challenges and partitions the 150 Northwind-Freeform pairs into training, development, and testing sets, balancing across participants' age, gender and depression levels.

We chose sixteen videos (see TABLE I.) in the testing set from the Freeform category with similar durations, ranging from 36s to 50s (Average = 41.2, Standard Deviation = 3.8), evenly across four depression categories.

B. Participants

Fourteen students with no prior knowledge of depression recognition were recruited to watch German language depression videos, see §D, below. Ethics Approval was obtained from the Australian National University Human Research Ethics Committee. No participants understood German. Two were excluded due to technical failures of the sensors. The final sample was twelve participants, six males and six females, from 18 to 27 years in age (Average = 21.1, Standard Deviation = 2.8) with normal or corrected-to-normal vision and hearing. This sample size of participants is normal for publications of a preliminary study in medicine [18].

C. Measures and Sensors

1) Galvanic Skin Response (GSR): GSR, also known as skin conductance (SC) or electrodermal activity responses (EDA), measures an individual's electricity flow through the skin, which varies due to the amount of sweat on the skin [19]. The GSR is composed of two separate electrodermal activities. The tonic component is a slow-moving signal that shows the general activity of the perspiratory glands caused by body or external temperature, while the phasic component is the faster distinctive waveform in the signal, and is considered to be linearly correlated with the intensity of arousal in mental state [20]. In this study, GSR was recorded by Empatica E4 wristband with a sampling rate of 4Hz [21].

2) Skin Temperature (ST): ST fluctuates due to vasodilatation of peripheral blood vessels induced by increased activity of the sympathetic nervous system. It has been found to be negatively correlated with unpleasant emotions such as stress and fear [13] because blood is redirected to vital organs as protection measure. In this study, wrist ST was recorded using Empatica E4 wristband with a sampling rate of 4 Hz [21].

3) Pupillary Dilation (PD): PD provides indications of changes in mental states and the strengths of mental activities [22]. Pupil size was found to constitute a response to emotionally engaging stimuli where pupil is significantly bigger after positively and negatively arousing stimuli than after neutral stimuli [22]. In this study PD was recorded using EyeTribe eye tracker with a sampling rate of 60 Hz [23].

D. Procedure

A schematic diagram of the equipment setup is shown in Fig. 1. The experiment was conducted with each individual participant in the same quiet experiment room. Each participant was given a written set of experiment instructions and guidance from the experiment instructor before they provided written informed consent. Afterwards, the Empatica E4 sensor [21] was attached to the wrist of participant's non-dominant hand and the eye gaze calibration for the eye tracker was performed. Participants then filled in a questionnaire to collect demographic and health characteristics that may affect cardiovascular and pupillary responses. Each participant then watched 16 videos and were asked by the end of each video to respond to a question of "How would you like to rank the patient's depression level?" on a four-item scale of "None, Mild, Moderate, Severe" that matches with the BDI-II [5] scale. A five second gap was provided between every two videos. The videos were presented in an order balanced way to avoid effects of presentation ordering. After finishing watching all videos, participants also did the BDI-II [5] survey accessing their own depression level. In total, the experiment took approximately forty minutes.

After data collection, 60 observations from participants watching individuals with each depression category (none, mild, moderate and severe) were obtained. This resulted in 240 complete responses, including participants' conscious depression judgements and physiological sensors recordings.

TABLE I. STIMULI VIDEOS SELECTED FROM THE TESTING SET OF THE FREEFORM TASK IN AVEC 2014

Video name	Duration (in seconds)	Depression level	Depression category
210_2_Northwind_video	43	1	no
249_1_Northwind_video	42	4	no
341_1_Northwind_video	39	7	no
240_3_Northwind_video	41	11	no
220_3_Northwind_video	39	15	mild
242_1_Northwind_video	42	16	mild
315_3_Northwind_video	40	17	mild
214_3_Northwind_video	43	18	mild
245_3_Northwind_video	40	21	moderate
218_3_Northwind_video	39	22	moderate
325_2_Northwind_video	39	25	moderate
250_1_Northwind_video	41	27	moderate
226_2_Northwind_video	41	30	severe
359_1_Northwind_video	45	33	severe
315_2_Northwind_video	58	34	severe
237_1_Northwind_video	47	41	severe

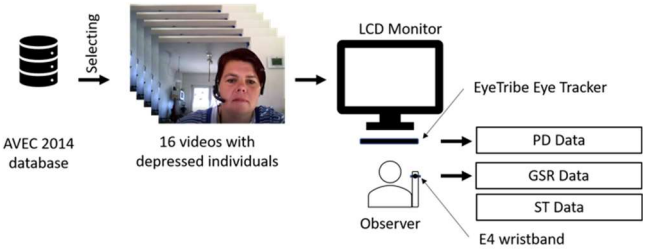

Fig. 1. A schematic diagram of the equipment setup.

III. METHODOLOGY

A. Preprocessing

Transient noise was observed in the raw physiological signals due to the movement of participants, which mostly happened at the beginning and the end of the recording when they were filling in the demographic questionnaire and post-experiment survey. Thus, for all participants and all signals, we first extracted the raw signal data when participants were watching the full set of 16 videos. Cubic spline interpolation was then applied to construct missing pupil size data caused by occasional eye blinks. This procedure was employed on the pupil data of left and right eyes separately.

Physiological signals are individual-dependent, meaning that different individuals may have the same physiological signal in different ranges. To reduce the between-participant differences, normalisation methods are needed and for this study, we applied a min-max scaler to all physiological signals separately, which scaled signals to a range between 0 and 1:

After normalisation, we smoothed the signals to remove noise artefacts. For GSR and ST, we used a lowpass Butterworth filter with an order of 6 and a cut-off frequency of 0.2 Hz and 0.3 Hz respectively [20], [24] to form a low-passed (LP) GSR and ST signal. For PD, we applied a 10-point Hann moving window average to left pupil and right pupil data separately [25].

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 invoked by his or her experience of watching one video.

B. Features Extraction

Once the data were pre-processed, it was necessary to extract features from the signals to characterise physiological patterns for different depression watching experiences. From the three physiological signals, we extracted a total of 85 time domain features for each video watching session, mainly focusing on capturing the amplitude variance and the occurrences of transient changes in the signals.

1) GSR features: According to the literature on using physiological signals for emotion recognition [25], we extracted a total of 23 features from normalised and filtered GSR signals. The following 8 features were extracted from both the normalised and filtered GSR signals separately.

a) Minimum

b) Maximum

c) Mean

d) Standard Deviation

e) Variance

f) Root Mean Square

g) Means of the absolute values of the first difference

h) Means of the absolute values of the second difference

As indicated in [20], [26], GSR contains two types of electrodermal activity: the slow-moving tonic component (also called DC level) that reflects the general activity of the perspiratory glands caused by body or external temperature, and the faster phasic component (also called the skin conductance response (SCR)) as a distinctive waveform in the signal that is considered to be linearly correlated with the intensity of arousal in mental state. To extract the DC level component, we applied a very low pass Butterworth filter with a cut-off frequency of 0.08 Hz on the normalised GSR signal to form the Very Low Pass signal (VLP). Since the LP GSR signal obtained from the low pass filter during pre-processing contains both DC level component and SCR component, to further acquire a detrended SCR data without DC component, we removed the continuous piecewise linear trend in both LP and VLP signal. Then we calculated the number of SCR occurrences for VLP, LP and normalised GSR signal separately, the mean of amplitudes of all these occurrences, and the ratio of SCR occurrences in VLP to the occurrences in LP, which form the following seven features.

i) Numbers of SCR occurences for VLP, LP and normalised signal

j) Amplitudes of SCR occurrences for VLP, LP and normalised signal

k) Ratio of SCR occurrences in VLP to ovvurrences in LP

2) PD features: We extracted features similar to those of the GSR signal. 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, means of the absolute values of the first and second difference were calculated as features. A very low pass Butterworth filter with a cut-off frequency of 0.08 Hz was applied 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 averge VLP PD signal. 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 were also extracted as features.

3) ST features: Similar features to those of the GSR signal were calculated. The minimum, maximum, mean, standard deviation, variance, root mean square, means of the absolute values of the first and second difference were calculated as features from the normalised and LP ST signal. A very low pass Butterworth filter with a cut-off frequency of 0.08 Hz was applied to the normalised ST signal to form the VLP ST signal. Numbers and amplitudes of peak occurences for VLP and LP ST signals as well as the ratio of peak occurances in VLP to those in LP were also extracted.

In the end, we collected a total of 85 features from the three physiological signals: 23 (GSR) + 39 (PD) + 23 (ST).

C. Features Selection

Large numbers of features can be derived from physiological signals to make predictions. However, this full set of features may include redundant and irrelevant features which may outweigh the more effective features, affecting classification performance. Also training a classifier with a large number of features can be computationally expensive. Hence, selecting effective features is critical, and has been known to improve quality of pattern recognition because it can assist the model in better capturing necessary patterns [13].

We used Genetic Algorithm (GA) to select better subsets of features as candidate chromosomes by determining the presence (1) or absence (0) of every possible feature in the model, based on the generalization performance of a classifier as the fitness function. The initial population for the

GA was set to use all features. A chromosome was set as a binary string where index for a bit represented a feature and the bit value indicated whether the feature was used for classification. An example of such representation is demonstrated in Fig. 2. All settings for GA used in the hybrid classification system can be found in TABLE II.

D. Neural Networks Based Classification Models

In this paper, we built two Neural Network (NN) based depression classification models:

- NN: a NN classification model that used all depression features as input to recognise depression patterns
- GA+NN: a NN that used a subset of features selected by a GA to recognise depression patterns

All NNs were fully connected neural networks with a sigmoid hidden layer of size 50 and an output layer of four output neurons, representing the four depression levels. The number of hidden neurons was set to 50 after we tested our neural networks with different hidden neuron size from 10 to 100 and found 50 the optimal for our task. All NNs were trained with the Adam optimizer [27] using backpropagation with the Cross-Entropy loss function.

The most common method for cross-validation is k-fold which randomly partitions data into k equal sized subsamples and uses one subsample as validation data, and the remaining k-1 as training data. However, for human data, a continuous segment of physiological data with more than one data point can reflect a human's responses to a stimulus. Training a classifier on random splits of data is not adequate, unless all data from one human is guaranteed to be within either the training set or the testing set for each run. This method of leaving all data for one human out is called leave-one-participant-out, which was used in this study. Physiological data from one observer was used as the testing set, and those from the remaining participants formed the training set, and repeated for all, averaging to get the final results.

In this study we were interested in recognising individuals' depression levels achieved by a combination of GSR, ST, and PD measurements as monitored signals. Assessment of the overall usefulness of each signal is also important as fewer sensors are required if only a single signal is needed to achieve

	Presence Vector of best representative features selected by GA	**1**	0	**1**	…
×	Vector of derived representative features	**0.2**	0.3	**0.5**	…
	Vector for best representative features selected by GA	**0.2**	0	**0.5**	…

Fig. 2. GA representation of features.

TABLE II. IMPLEMENTATION SETTINGS FOR GA

GA Parameter	Value
Population size	100
Crossover rate	0.8
Mutation rate	1/(length of chromosome)
Crossover type	Uniform crossover
Mutation type	Uniform mutation
Selection type	Stochastic uniform selection

classification comparison using NN and GA+NN under four conditions: 1) GSR+ST+PD: using all features extracted from all three signals; 2) GSR: only using features from the GSR signal; 3) ST: only using features from the ST signal; and 4) PD: only using features from the PD signal. It is worth noting that when each physiological was used, the classifier was retrained and retest using the same validation scheme.

E. Evaluation Measures

To validate the effectiveness of our models, we used *precision, recall* and *F1-score* as evaluation measures. For a specific depression level L, *precision* is defined as the proportion of individuals that are correctly predicted with depression level L actually have the depression; *recall* is the percentage of depressed individuals that are correctly predicted with depression level L among all individuals labelled with depression level L; and *F1 score* takes the harmonic mean of precision and recall defined as $2 \times (Predicion_L \times Recall_L)/(Precision_L + Recall_L)$.

As multiclass depression labels were predicted, we calculated the average precision, recall and F1 score for all depression levels to give a view on the general prediction performance. We also computed the overall accuracy to evaluate the overall performance, which is the number of individuals correctly predicted with their corresponding depression levels over the total number of individuals.

IV. Results

A. Observers Subjective Prediction

As can be seen from TABLE III. , our observers were not very good at consciously identifying the depression level of individuals in videos. The overall accuracy was 27% which is slightly over the prima-facie chance level of 25% since there were four options for observers over balanced numbers of video stimuli. The average ratios of consciously identifying healthy individuals and severely depressed individuals correctly were 33% and 29% respectively, higher than those of identifying depressed individuals in middle ranges, at 21% and 25%. This could imply that people are better at identifying healthy individuals, and depressed patients with high severity, but worse at differentiating depression levels.

B. Classification based on All Physiological Signals

Two classification models were tested on the physiological data obtained from the depression experiment. All features derived from observers' GSR, ST, and PD signals were provided to the NN and GA+NN models. Performances of the classifications were calculated based on the average results of 10 runs, and the results are shown in TABLE IV. .

Signal patterns in observers viewing the videos of individuals with varying levels of depression were better recognised with the GA hybrid according to the evaluation measures. When the NN was provided with all features derived from the physiological signals, the average precision, recall and the F1 score of four depression levels, and the overall accuracy across all levels were 4% lower than those of the model with GA feature selection, at a rate of 92%.

Similar to observers' subjective predictions, both GA and GA+NN were less accurate in predicting depression levels of individuals with mild and moderate depression severity. GA+NN outperformed NN in predicting "None", "Mild" and "Severe" level, achieved by having higher precision, recall and F1 score. However, for level "Moderate", GA+NN has

lower precision and F1 score, possibly meaning GA+NN compared to NN is more biased to this depression level and thus tends to predict more people with "Moderate" depression.

C. Classification based on a Single Physiological Signal

In order to evaluate the classification capability of models with fewer physiological signals, features derived from single physiological signal were provided to the NN and GA+NN models. Performances of the classifications were calculated based on the average results of 10 runs and the overall accuracies for seven conditions are shown in Fig. 3.

Among the three single physiological signals, models trained using ST features performed the worst compared to those using GSR and PD. GA improved the NN model trained using ST features from an overall accuracy of 84% to 87%, however, it did not create a difference for models trained with GSR and PD. Both NN and GA+NN models trained with GSR features achieved an overall accuracy of 89% while PD features contributed to more accurate models with an accuracy of 92%. These could imply that observers' PD itself can be an effective signal for predicting other individuals' depression level while GSR and ST may provide less informative features.

When GA is not used for feature selection, an NN model trained with features from all signals performed slightly worse than that trained with PD signal. This shows that the NN is susceptible to features that were irrelevant and redundant for depression classification. The better performance of GA+NN trained with all signals confirmed this.

To further see the contribution of each signal to the prediction of each depression level, precision, recall and F1 score of each depression level were calculated for both NN and GA+NN. The results are shown in **Error! Reference source not found.**

PD was the best in recognising all depression levels except the "None" category, achieved by having the highest precision, recall and F1 score for both NN and GA+NN. GSR gave the best result in identifying individuals with no depression levels indicated by its highest performances in all three measures for NN and GA+NN. This may indicate when only one signal is available, GSR is more useful in identifying healthy people among the depressed individuals while PD is better at differentiating depressed people by categories.

V. Discussion

With this preliminary study, we explored people's automatic and non-conscious physiological responses to observing individuals with four depression severities, as well as their more conscious subjective judgments of the depression level in a subject's video. For people's ability to accurately identify the depression level of other individuals, conscious judgment would not detect the correct level much better than chance. Consistent with earlier findings about the accuracy of people's conscious judgments on the veracity of smiles [14], anger [15], and dishonesty [16], observers in this study were found to be accurate about at a chance level. This could imply that human conscious judgments on other individuals' depression severity may be slightly above chance in general. Future research could explore the accuracy of conscious judgments from trained medical personnel who are skilled at diagnosing depression patients.

Observers' GSR, ST and PD were measured during the viewing of individuals with different depression levels. Models trained with features from these three signals yielded acceptable results (up to an overall accuracy of 92% accuracy across all levels) of identifying four depression levels, measured by the patients using the Beck II schedule, which were much higher than observers' subjective judgments. The results provide evidence that although humans cannot consciously recognise the depression severity of other individuals correctly, and have a better ability of identifying depression at an unconscious level and this ability can be accessed by computational classifiers and sensors. The superior depression detecting ability of human unconscious physiological responses over conscious judgments also occurs

TABLE III. Results of Depression Prediction from Observers' Verbal Responses

Depression level	Subjective Prediction		
	Precision	*Recall*	*F1 score*
None	0.31	0.33	0.32
Mild	0.18	0.21	0.19
Moderate	0.23	0.25	0.24
Severe	0.42	0.29	0.35
Average	0.29	0.27	0.28
Overall Accuracy	0.27		

TABLE IV. Performance Measures for Depression Recognition Models Defined from All Physiological Signals

Depression level	NN			GA+NN		
	Precision	*Recall*	*F1 score*	*Precision*	*Recall*	*F1 score*
None	0.85	0.90	0.87	0.92	0.95	0.94
Mild	0.85	0.85	0.85	0.93	0.89	0.91
Moderate	0.92	0.88	0.90	0.88	0.90	0.89
Severe	0.91	0.89	0.90	0.95	0.95	0.95
Average	0.88	0.88	0.88	0.92	0.92	0.92
Overall Accuracy	0.88			0.92		

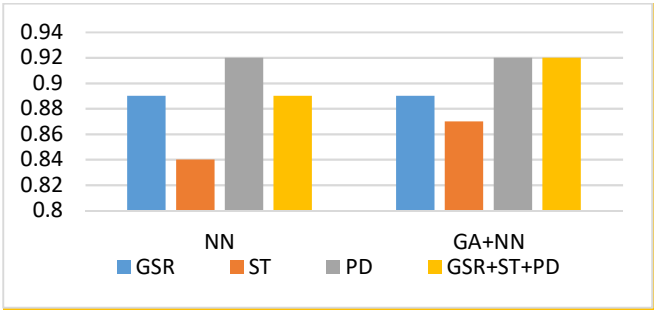

Fig. 3. Overal Accuracies of Depression Prediction from Single Physiological Signal and All Signals.

TABLE V. Performance Measures by Depression Level for Models Defined from Single Physiological Signal

Depression level	Signal	Precision		Recall		F1 score	
		NN	*GA+NN*	NN	*GA+NN*	*NN*	*GA+NN*
None	GSR	**0.94**	**0.94**	**0.91**	**0.93**	**0.92**	**0.93**
	ST	0.87	0.85	0.88	0.9	0.87	0.87
	PD	0.89	0.88	**0.91**	**0.93**	0.9	0.9
Mild	GSR	0.86	0.85	0.84	0.84	0.85	0.85
	ST	0.86	0.88	0.83	0.86	0.84	0.87
	PD	**0.92**	**0.93**	**0.91**	**0.9**	**0.92**	**0.92**
Moderate	GSR	0.85	0.88	0.88	0.87	0.87	0.88
	ST	0.79	0.84	0.84	0.86	0.81	0.85
	PD	**0.94**	**0.92**	**0.94**	**0.94**	**0.94**	**0.93**
Severe	GSR	0.91	0.88	0.94	0.9	0.92	0.89
	ST	0.86	0.9	0.83	0.85	0.85	0.88
	PD	**0.92**	**0.95**	**0.91**	**0.92**	**0.92**	**0.94**

in other areas such as estimating realness of basic emotions [14], [15]. This suggests that unconscious responses from human instinctive ability, which has been adaptively evolved by natural selection, can make efficient and effective use of cues of identifying depressed individuals despite of the influence from conscious biases.

Additionally, the results of classification with features from all available physiological signals and with features from fewer signals reveal that some physiological signals seem to convey more information to the classifier. In this study, models trained with PD-only performed the best classification among models trained with other single signals, achieving the highest overall accuracy and the highest precision, recall and F1 score for Mild, Moderate and Severe depression level. This phenomenon is consistent with the literature [28] where pupil size was found to be prominent among other signals in detecting stress. On the other hand, models trained with GSR-only achieved the highest measure in identifying healthy individuals. This may indicate that PD could be an effective indicator of other individuals' depression state and GSR is useful in identifying a healthy state.

VI. Limitations and Future Work

Observers in this study are naive individuals who do not have any experience of diagnosing depression. They do not understand German which is the language spoken by the individual in the video. In a future study, more psychologists skilled in diagnosing depression and German speakers should be recruited to examine the effect of domain knowledge and language understanding on depression prediction. Stronger conclusions may also be able to be drawn in subsequent studies with more observers. Finally, observers' other physiological signals could also be investigated.

VII. Conclusion

Our work explored the use of physiological signals from observers to detect depression level of individuals in videos. When individuals with different depression severities were observed, three physiological signals, GSR, ST and PD, were affected. After pre-processing, these signals generated a total of 85 features which could be used to train neural networks to predict other people's depression level with an accuracy up to 92%. This accuracy did not drop when only PD was used with GA as feature selection. We demonstrated that neural networks trained with observers' physiological signals are powerful indicators of other individuals' depression severity. Future research and implementation of the findings in this area are likely to be beneficial in assisting with more objective and earlier depression diagnosis, which combined with the use of known effective treatments would be beneficial to the society.

References

[1] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," Speech Commun., vol. 71, pp. 10–49, 2015.

[2] K. Hawton, C. C. i Comabella, C. Haw, and K. Saunders, "Risk factors for suicide in individuals with depression: a systematic review," J. Affect. Disord., vol. 147, no. 1–3, pp. 17–28, 2013.

[3] J. J. Mann et al., "Suicide prevention strategies: a systematic review," Jama, vol. 294, no. 16, pp. 2064–2074, 2005.

[4] A. J. Mitchell, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: a meta-analysis," Lancet, vol. 374, no. 9690, pp. 609–619, 2009.

[5] A. T. Beck, R. A. Steer, and G. K. Brown, "Beck depression inventory-II," San Antonio, vol. 78, no. 2, pp. 490–498, 1996.

[6] M. Hamilton, "A rating scale for depression," J. Neurol. Neurosurg. Psychiatry, vol. 23, no. 1, p. 56, 1960.

[7] J. Joshi et al., "Multimodal assistive technologies for depression diagnosis and monitoring," J. Multimodal User Interfaces, vol. 7, no. 3, pp. 217–228, 2013.

[8] S. Potvin, G. Charbonneau, R.-P. Juster, S. Purdon, and S. V. Tourjman, "Self-evaluation and objective assessment of cognition in major depression and attention deficit disorder: Implications for clinical practice," Compr. Psychiatry, vol. 70, pp. 53–64, 2016.

[9] J. Tomaka, J. Blascovich, R. M. Kelsey, and C. L. Leitten, "Subjective, physiological, and behavioral effects of threat and challenge appraisal.," J. Pers. Soc. Psychol., vol. 65, no. 2, p. 248, 1993.

[10] S. Scherer et al., "Automatic behavior descriptors for psychological disorder analysis," in Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, 2013, pp. 1–8.

[11] Y.-T. Chen, I.-C. Hung, M.-W. Huang, C.-J. Hou, and K.-S. Cheng, "Physiological signal analysis for patients with depression," in Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on, 2011, vol. 2, pp. 805–808.

[12] F. A. Jain et al., "Heart rate variability and treatment outcome in major depression: a pilot study," Int. J. Psychophysiol., vol. 93, no. 2, pp. 204–210, 2014.

[13] N. Sharma and T. Gedeon, "Modeling stress recognition in typical virtual environments," in Proceedings of the 7th international conference on pervasive computing technologies for healthcare, 2013, pp. 17–24.

[14] M. Z. Hossain, T. Gedeon, and R. Sankaranarayana, "Observer's galvanic skin response for discriminating real from fake smiles," Australas. Conf. Inf. Syst., pp. 1–8, 2016.

[15] 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.

[16] X. Zhu, Z. Qin, T. Gedeon, R. Jones, M. Hossain, and S. Caldwell, "Detecting the Doubt Effect and Subjective Beliefs Using Neural Networks and Observers' Pupillary Responses: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13-16, 2018, Proceedings, Part IV." pp. 610–621, 2018.

[17] M. Valstar et al., "Avec 2014: 3d dimensional affect and depression recognition challenge," in Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, 2014, pp. 3–10.

[18] R. Simon, "Optimal two-stage designs for phase II clinical trials," Contemp. Clin. Trials, vol. 10, no. 1, pp. 1–10, 1989.

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

[20] 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.

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

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

[23] TheEyeTribe, "The EyeTribe." [Online]. Available: http://theeyetribe.com/theeyetribe.com/about/index.html. [Accessed: 30-May-2018].

[24] V. Xia, N. Jaques, S. Taylor, S. Fedor, and R. Picard, "Active learning for electrodermal activity classification," in Signal Processing in Medicine and Biology Symposium (SPMB), 2015 IEEE, 2015, pp. 1–6.

[25] 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.

[26] M. Züger and T. Fritz, "Interruptibility of software developers and its prediction using psycho-physiological sensors," in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015, pp. 2981–2990.

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

[28] J. Zhai and A. Barreto, "Stress detection in computer users through non-invasive monitoring of physiological signals," Blood, vol. 5, no. 0, 2008.