Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors

Sharifa Alghowinem, Roland Goecke, *Member, IEEE*, Michael Wagner, *Senior Member, IEEE*, Julien Epps, *Member, IEEE*, Matthew Hyett, Gordon Parker, and Michael Breakspear

Abstract—An estimated 350 million people worldwide are affected by depression. Using affective sensing technology, our *long-term goal* is to develop an objective multimodal system that augments clinical opinion during the diagnosis and monitoring of clinical depression. This paper steps towards developing a classification system-oriented approach, where feature selection, classification and fusion-based experiments are conducted to infer which types of behaviour (verbal and nonverbal) and behaviour combinations can best discriminate between depression and non-depression.Using statistical features extracted from speaking behaviour, eye activity, and head pose, we characterise the behaviour associated with major depression and examine the performance of the classification of individual modalities and when fused. Using a real-world, clinically validated dataset of 30 severely depressed patients and 30 healthy control subjects, a Support Vector Machine is used for classification with several feature selection techniques. Given the statistical nature of the extracted features, feature selection based on T-tests performed better than other methods. Individual modality classification results were considerably higher than chance level (83 percent for speech, 73 percent for eye, and 63 percent for head). Fusing all modalities shows a remarkable improvement compared to unimodal systems, which demonstrates the complementary nature of the modalities. Among the different fusion approaches used here, feature fusion performed best with up to 88 percent average accuracy. We believe that is due to the compatible nature of the extracted statistical features.

Index Terms—Depression detection, multimodal fusion, speaking behaviour, eye activity, head pose

1 Introduction

Fluctuations in mood are a normal part of most people's emotional lives, as long as such fluctuations are not severe, frequent, or interfere with that individual's daily and social life functions. If they do, a psychiatric disorder such as major depression disorder might be present. Major depression is a mood disorder that may last for weeks, months, even years, vary in severity, and is associated with distress and disability that impair an individual's ability to function in daily life. The World Health Organisation (WHO) lists depression as the fourth most significant cause of disability worldwide and predicts it to be the leading cause in 2020 [1]. Moreover, in a recent report, the WHO estimated that 350 million people worldwide are affected by depression [1]. The suicide risk is more than 30 times higher among depressed than in the general population [2].

Treatment of depression disorders is effective in many cases [3], but misdiagnosing depressed patients is a common barrier [4]. Although clinical depression is one of the most common mental disorders, it is often difficult to diagnose, because it manifests itself in different ways and because clinical opinion and self-assessment are currently the only means of diagnosis, risking a range of subjective biases. According to the WHO Global Burden of Disease report, the barriers to effective diagnosis of depression include a lack of resources and trained health care providers. Moreover, evaluations by clinicians vary depending on their expertise and the diagnostic methods used (e.g., Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [5]). Depression has no dedicated laboratory tests and hence, there is currently no objective method to diagnose depression. We believe that recent developments in affective sensing technology will potentially enable an objective assessment. While automatic affective state recognition has been an active research area in the past decade, methods for mood disorder detection, such as depression, are still in their infancy.

Our *ultimate goal* is to develop an objective multimodal affective sensing system that supports clinicians during the diagnosis and monitoring of clinical depression. In the *long*

- *S. Alghowinem is with Prince Sultan University, Riyadh 11586, Saudi Arabia. E-mail: sharifa.alghowinem@anu.edu.au.*
- *R. Goecke is with the University of Canberra, Canberra ACT 2617, Australia. E-mail: roland.goecke@ieee.org.*
- *M. Wagner is with the University of Canberra, Canberra ACT 2617, Australia, the Australian National University, Canberra ACT 0200, Australia, the National Centre for Biometric Studies Pty Ltd, Canberra ACT 2600, Australia, and with the Technical University of Berlin, 10623, Berlin, Germany. E-mail: michael.wagner@canberra.edu.au.*
- *J. Epps, M. Hyett and G. Parker are with University of New South Wales, Sydney NSW 2052, Australia. E-mail: {j.epps, m.hyett}@unsw.edu.au, g.parker@blackdog.org.au.*
- *M. Breakspear is with QIMR Berghofer Medical Research Institute, Brisbane QLD 400, Australia and the Metro North Mental Health Service Brisbane, QLD 4029, Australia. E-mail: mjbreaks@gmail.com.*

Manuscript received 29 Jan. 2015; revised 12 Oct. 2016; accepted 23 Nov. 2016. Date of publication 30 Nov. 2016; date of current version 5 Dec. 2018. Recommended for acceptance by L.-P. Morency.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2016.2634527

term, such a system may also become a very useful tool for remote depression monitoring to be used for doctor-patient communication in the context of e-health infrastructure. Clinical assessment of patients with depression relies heavily on two domains—the clinical history (i.e., history of presenting symptoms, prior episodes, family history etc.) and the mental state examination (appearance, speech, movement, reported mood etc., i.e., the assessment of affect). The latter is the focus here. In particular, we investigate the analysis of audio-visual data acquired during a clinical interview with people meeting criteria for depression for features that would generally be assessed for the classic mental state examination. Nevertheless, this is not to imply that this investigation of behavioral analysis could replace the mental status examination, but rather to supplement it.

The aims of this paper are:

- Investigating behavioural characteristics of depressed subjects compared with matched controls using speech style, eye activity and head pose modalities individually. This is done by exploring robust and quantitative differences between the depressed group and control group that are not prone to the subjective biases outlined above. This investigation is performed to aid in selecting the most promising behavioural features for the multimodal classification.
- Exploring several data fusion techniques to capture multimodal patterns of behavioural changes, following the holistic (integrative) approach of the clinician.

This paper examines possibly the broadest array of behavioural features to date and studies these empirically on a clinically annotated matched control-depressed database that is suitable for multimodal behavioural analysis. This is significant because behaviour has long been closely associated with depression-related symptoms, and offers a promising non-invasive possibility for automatic depression assessment. These behavioural features are statistically analysed for the feature selection step before the classification. Moreover, this research's *long-term goal* is to develop an auotomated multimodal depression detection system in a classification framework.

This paper is also the first to undertake a detailed and comprehensive empirical study of methods for fusing depression-related behavioural indicators (based on statistical measures) from different sensor modalities for depression detection. This is significant because depression is a complex and multi-factor disorder and to date it appears very likely that information from multiple modalities will be essential in addressing automatic depression assessment.

2 Background and Related Work

In the last two decades, affective sensing has been an active research area, used in many contexts. One particular application area in recent years has been in automated depression analysis. It seems obvious that a multimodal system that fuses different channels and cues is expected to provide more accurate recognition than unimodal approaches. However, only few affective sensing systems use multimodal input where different modalities are fused, such as body movement, facial expression and speech prosody, as reviewed in [6], [7]. Moreover, the AVEC depression challenges have attracted much interest lately for assessing systems to predict the depression severity [8]. Yet, relatively few systems employed a multimodal approach as reviewed below.

D'Mello and Kory [9] analysed some of these studies by comparing the unimodal with multimodal results in the affect detection domain. Regardless of the considerable variation in terms of data, affect, modality, and method, a consistent improvement was found for the multimodal approach [9]. However, the fusion of different modalities is not a trivial task. Several issues of when and how to combine those modalities have to be considered. For example, fusion could be performed as pre-matching (early) fusion and post-matching (late) fusion.

Early fusion can be executed on raw data from each sensor (sensor fusion) or on the extracted features from raw data (feature fusion). Even though early fusion is expected to contain richer information than late fusion [10], it comes with complications and weaknesses. For example, feature vectors from different modalities might be unaligned, and incompatible. Such vectors could have different time scales (i.e., different duration) or sampling rates, for example, combining speech with video when there are silence periods. Differences in dimensionalities or sizes could be problematic in early fusion, such that combining one-dimensional data (e.g., speech) with two-dimensional data (e.g., video) could introduce a bias within the classifier for one modality or another.

Several techniques for sensor fusion have been investigated to overcome the above weaknesses and, therefore, increase the robustness of the fusion results as reviewed in [11], [12], [13]. These studies showed that measures of reliability and/or confidence are required for robust sensor fusion using computational theory.

Incompatibility issues have to be remedied before fusing features using normalisation methods such as min-max, Z-score, etc. Once normalised, features can be simply concatenated or pre-processed for dimensionality reduction. Reducing dimensionality can be performed by feature selection or feature transformation. Feature selection is a statistical technique to find a relevant subset of features from original features, for example, using statistical search techniques to determine the most promising subset of features. In general, feature selection methods can be divided into three categories: filters, wrappers, and embedded methods [14]. Wrappers and embedded approaches utilise classification techniques to select the feature subset, which can risk overfitting issues especially for small datasets. On the other hand, filters select a subsets of features, independently of any classification algorithms, using statistical measures such as ranking, correlation or simple tests methods. Feature transformation creates new features using functions of the original features such as principal component analysis (PCA) [15].

In this work, sensor fusion using raw data is not investigated, not only due to incompatibility in sampling rate, segmentation, etc. between audio and video channels, but due to unsuitablilty for investigating behavioural characteristics of depression. For example, silence segments in the audio signal were eliminated, while the same segments on video signal were used. For early fusion investigated in this work, extracted features from the sensor data were normalised before fusion.

Late fusion is executed after the classification of each individual channel, using either the classifier output scores (score fusion) or labels (decision fusion). Fusing scores from different modalities that use the same type of classifier is simple. However, fusing scores from different types of classifiers can be complicated if the scores are not compatible (i.e., distance from hyperplane versus likelihood ratio), therefore further normalisation before the fusion should be performed [16]. Both score and decision fusion could be executed in simple ways (e.g., sum-rule, product-rule, etc., and logical AND, majority voting, etc.), or in more complex ways such as using a secondary classifier [16].

Hybrid fusion was recently introduced to utilise benefits of both early and late fusion, i.e., fusing the scores or decisions from both feature-concatenated and individual modalities [17]. Hybrid fusion can be implemented by having either one or two levels of score/decision fusion. Either way, feature fusion of all modalities is performed first to create a new modality, which is then treated as an additional individual modality. The scores/decisions of this new modality are then fused with the scores/decisions of the individual modalities in either one or two levels (see Fig. 3d).

To explore the most suitable fusion technique for our multimodal depression classification investigation, this work investigates both late and hybrid fusion methods using several approaches. Noting that advanced approaches of late and hybrid fusion that require separate sets for training, testing and evaluation data could not be investigated due to the relatively small dataset used here.

Many previous studies on automatic detection of depression have only investigated a single channel, either from video or audio. To the best of our knowledge, only a few studies have investigated multiple channels for this task [18], [19], [20], [21], [22], [23]. In [18], the relationship between facial actions and vocal prosody for clinical depression detection was explored. However, the study did not investigate fusion approaches for the examined channels. Scherer et al. [19] investigated audio-visual indicators for automatic depression detection, which were concatenated using feature fusion. The fused modalities outperformed individual ones significantly, resulting in 90 percent accuracy (compared to 51 and 64 percent for acoustic and visual modalities alone, respectively).

In [20], the GMM-UBM system for the audio subsystem and Space Time Interest Points in a Bag-of-Words approach for the vision subsystem were fused at feature level. Even though the improvement in the fused system was not statistically significant from the individual subsystems in detecting the depression severity, other fusion approaches were not investigated. Recently, Williamson et al. [21] correlated several speech features along with facial action unit features with the severity of depression in a multimodal system using score fusion method, which achieved good results to predict depression severity. Meng et al. [22] also investigated fusing facial and vocal expressions for this task, fused using a weighted sum decision fusion, and the result improved slightly from individual channels.

In our previous work [23], several multimodal (audio-video) fusion techniques using only low-level features were compared at feature level, score level and decision level for depression analysis. The low-level features were clustered using Bag-of-Audio features for the audio channel and Bag-of-Video features for Space Time Interest Points for the video channel and then analysed individually and combined for detecting depression, showing considerable improvements in the fused system compared with individual modalities.

As can be noticed, the multimodal investigation in the previously mentioned depression detection studies is not only limited in fusion approaches (e.g., feature fusion, score fusion), but also for the number and type of explored modalities (i.e., speech and facial only). Therefore, in this paper, to evaluate a multimodal approach for the automatic detection of depression, we investigate several fusion techniques for classifying depression characteristics from speech behaviour, eye activities, and head pose and compare the results with the unimodal results.

3 Method

3.1 Participants and Data Acquisition

Clinically validated data was collected at the Black Dog Institute[1]—a clinical research facility offering specialist expertise in depression and bipolar disorder—in Sydney, Australia. The study used healthy controls and subjects diagnosed with depression (either Melancholia or Major Depression Disorder (MDD)). We acknowledge the risk of treating Melancholia and MDD patients as one class here, however, a further division is not practical for the classification task, given the relatively modest sample size.

Depressed patients were recruited into the study from the tertiary referral Depression Clinic at the Black Dog Institute. All patients were classified as having a current major depressive episode on the Mini International Neuropsychiatric Interview (MINI [24]), conducted by trained research assistants (RA), with the type of depression (variably melancholic, non-melancholic and bipolar depression) rated independently by clinical psychiatrists. Healthy control subjects were recruited from the community. Exclusion criteria for healthy controls included current and/or past depression, (hypo)mania or psychosis as assessed by the MINI. Clinical participants were excluded if they met criteria for current and/or past psychosis (unrelated to mood). Additional exclusion criteria, for all subjects, included current and/or past alcohol dependence, neurological disorder or history of significant brain injury, a Wechsler Test of Adult Reading [25] score below 80 and/or electroconvulsive therapy in the past six months. Depression severity was assessed using the Quick Inventory of Depressive Symptomatology Self Report (QIDS-SR [26]), with all clinical participants meeting at least a moderate level of depression severity. Informed consent was obtained from all participants and the study proceeded with approval from the local institutional Human Research Ethics committee in line with the guidelines for human research from the National Health and Medical Research Council.

The audio-video experimental paradigm contains several parts, including an interview with a clinician, where specific open ended questions were asked. Subjects were asked to describe events in their life that had aroused significant emotions. This item was designed to elicit spontaneous self-directed speech and related facial expressions, as well as

1. http://www.blackdoginstitute.org.au

overall body language. The content of affective situations including neutral situations such as routine activities, positive social events, such as births and weddings, and negative situations, such as bereavement or financial problems, were explored, with a particular focus on perceived mechanisms leading to depression. Open ended questioning was conducted by one of two trained RAs, in a typical clinician-patient interaction during an assessment. The RAs were not blinded to the depression status of the subjects. Moreover, some of the subjects had previous interaction with the RAs (mostly from the control subjects subset).

Video and audio streams were captured in QuickTime Pro (running on a 17" Apple Macbook Pro) using a high-resolution Pike F-100 FireWire camera (Allied Vision Tech.), and broadcast-quality (Sony) lapel microphone. The camera was positioned on a tripod behind the monitor that presented the stimuli, with the height of the camera adjusted for each participant to ensure optimal recording of facial features. The microphone was attached to the participant's lapel, at mid-chest level. During open ended questioning, the RA stood camera-left, behind the monitor (to the right of the participant). Audio was digitised at 44.1 kHz, and the video frame rate was set at 30 fps (frame per second). Both depressed and control subjects were recorded using the same facility (same room setting, hardware equipment, and software). All sessions were recorded at the Black Dog Institute during office hours (8 am-5 pm). Moreover, the recordings were collected over three years.

Matched-subject design is an important concept for studies in psychology, that aims for equating groups on some variables to reduce their effect on skewing the results. Generally, matched-subject design (also referred to as between-subject design) is preferred as it is sensitive to the effects of the independent variable, which increases the statistical power. To the best of our knowledge, the dataset used in this work is the only clinically annotated matched control-depressed database that is being used for automated multimodal behavioural analysis. In this study, the gender and age were matched in depressed and control groups to reduce the variability of gender bias and the age differences effect. In this study, a gender balanced subset of 30 depressed subjects (19 Melancholia patients, 10 MDD patients, and 1 Bipolar patient) and 30 controls was used (age range 21-75yr, $\mu 38 \pm 14$). Only native Australian English speaking participants were selected, to reduce the variability arising from different language acquisition. For depressed subjects, the level of depression was a selection criterion, with a mean of 19 points of the diagnoses using QIDS-SR (range 14-26 points, where 11-15 points refer to a "Moderate" level, 16-20 points to a "Severe" level, and ≥ 21 points to a "Very Severe" level).

In this paper, only the interview part of the paradigm was analysed, as it contains spontaneous interaction behaviour for both audio and video channels. The total duration of the recorded video-audio interviews is over 500 minutes (see Table 1). In addition, the interviews were manually labelled to extract pure subject speech,[2] as well as reciprocal speech to extract speech behaviour (see Section 3.2.1). The total pure speech duration is 290 minutes (see Table 1).

2. Where speech of other speakers, overlapped speech, as well as pauses, noise, laughs, etc. are segmented.

TABLE 1
Total, Average and Standard Deviation Duration (in Minutes) of Depressed and Control Subjects Speech in the Interview Part

Part	Depressed	Control	Total
Duration of full interviews:			
Total	309.2	199.8	509.0
Average	10.1	6.6	8.4
Standard deviation	$\pm$5.5	$\pm$2.0	$\pm$4.4
Duration of subjects' speech			
Total	183.2	107.7	290.9
Average	6.1	3.6	4.8
Standard deviation	$\pm$4.3	$\pm$1.5	$\pm$3.5

3.2 Feature Extraction

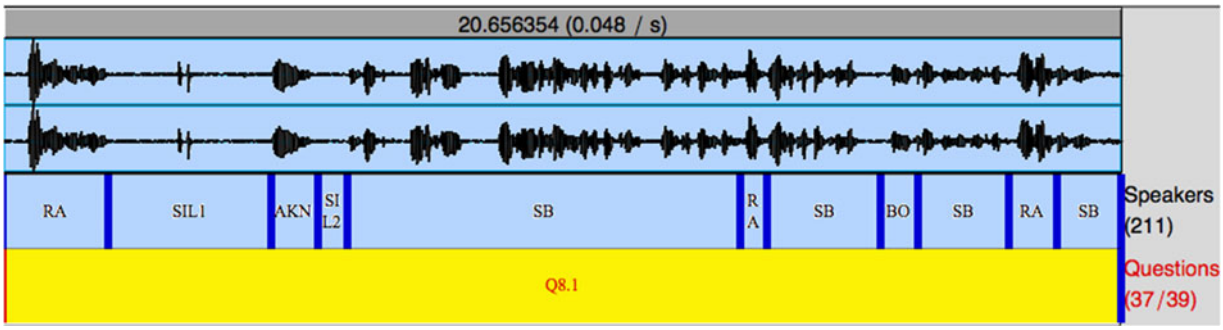
To analyse behavioural patterns of subjects' responses, we extracted statistical features from speech behaviour, eye activity, and head movement. As facial expressions to diagnose depression have been investigated in the literature, which mostly uses low-level features, and as the video recording did not include the full body to analyse body and hand movements, these modalities are not included in this paper, but are acknowledged as potential further relevant sources of information. Moreover, the focus here is to extract behavioural patterns (using statistical measures) of depressed subjects compared with healthy control subjects. For all modalities, some of the extracted features were selected based on the literature and the rest were hypothesised to be potentially relevant but had not been previously investigated for this task. Statistical analyses are carried out to filter out the insignificant features.

3.2.1 Speech Behaviour Features

Behaviourally based evaluations of depressed speech have found several distinguishing speech patterns as indicators of disease progression, severity or treatment efficacy [27], [28], [29]. In our preliminary study [30], we found that the response time and speech rate were longer in depressed subjects, while the interaction involvement and articulation rate were higher in control subjects. In previous work [30], [31], [32], [33], we intensively investigated vocal prosody features. In this paper, we focus on speaking behaviour patterns (speaking rate, pauses, and articulation rate) for the task of detecting depression and also for its compatibility with the statistical features extracted from other modalities as described below.

The subject interviews were manually labelled to separate speakers (i.e., interviewer and interviewee) and to separate the interview open-ended questions. Within each question, reciprocal speech and turns were also labelled (see Fig. 1). In this work, the speech behaviour feature extraction is divided into two parts: (1) extracting features from the extensive manual labelling, and (2) extracting speech rate features from subjects' segments.

Manually labelled speaker turns were used to extract several statistical measurements of the duration for analyses. A total of 63 statistical features are extracted from the manual labelling of the interview, grouped in seven duration groups:



RA: Research Assistant speech, SIL1: first silence lag, AKN: Acknowledgment, SIL2: second silence lag, SB: subject speech, BO: overlapped speech

Fig. 1. An example of manual labelling of interview speech: *The response lags, as well as pure and overlapped speech were labelled for feature extraction.*

- subject's speech,
- research assistant (RA) speech,
- time to first response, which is the duration of the silence after asking a question until an acknowledgement indicated by any sounds or words that are not the actual answer for the question (e.g., "ahhh"," hmm", "well", etc.),
- total response time, which is the lag between asking the question and the actual answer,
- subject laughing, which indicates a positive affective response in a conversation,
- overlapping laugh, and
- overlapping speech, which measures the involvement style in a conversation.

From each of the above duration feature groups, nine statistical features are calculated, namely: the average, maximum, minimum, range, variance, standard deviation, total, duration rate (duration of the feature in question ÷ total duration of the interview), and count (number of occurrences of the feature in question). This resulted in 7×9 features.

Speaking rate features were also extracted from subject speech segments by applying voice activity detector (VAD) using the Praat software. From the silent and sounding parts, speech, speaking, and pauses duration are extracted as listed in the following feature list. Moreover, to calculate speech rate and articulation rate, the number of syllables has to be calculated. We used a Praat script by [34], which calculates the number of syllables in a sounding segments. A further 19 speech style features are extracted as follows:

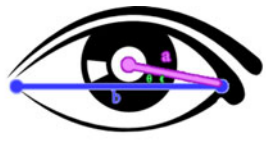
- Maximum, minimum, range, variance, and standard deviation, for sounding and silent parts (2×5 features),
- For sounding part:
  - number of sounding,
  - total speaking duration (excluding pauses),
  - articulation rate (number of syllable ÷ total speaking duration),
  - average speaking duration (speaking duration ÷ number of syllable),
- For silent part:
  - number of pauses,
  - total silence duration,
  - silence rate (number of pauses ÷ silence duration),
  - average silence duration (silence duration ÷ number of pauses),
- Number of syllables.

When measuring the speech rate, pauses are included in the duration time of the utterance, while the articulation rate excludes pauses [35]. Worth noting is that, unlike some other speech features (e.g., MFCCs), these are features that are (explicitly or implicitly) observed by clinical practitioners.

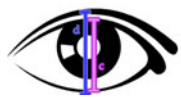
3.2.2 Eye Activity Features

Depressed patients were found to differ from the healthy comparison group in decreased direct eye contact with the interviewer, decreased eyebrow movement and elevated blink rates [28], [36], [37]. In our previous work [38], we found that the average distance between the eyelids while open was significantly smaller and the average duration of blinks significantly longer in depressed subjects.

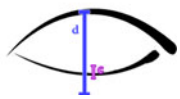
To accurately detect eye activities such as blinking and iris movements, the eyelids and iris need to be located and tracked. For this purpose, we trained and built special subject-specific 74-point eye active appearance models (AAM). To train the eye model, an average of 45 images per subject were manually selected having different eye status (e.g., open, half open, closed eye) and head position variation, then annotated. The annotated images were then used to build the eye model, using linear parameters to update the model in an iterative framework as a discriminative fitting method. For each eye, horizontal and vertical iris movement, and eyelid movement were extracted as low-level features per frame (30 fps) (see Fig. 2). It is worth noting that the right and left eyes are relative to the camera's view (for



(a) Left-right *(length of a ÷ length of b)* and up-down ($\theta = \cos^{-1}(\frac{a.b}{|a|.|b|})$) movement features



(b) Eyelid distance of open eye *(length of c ÷ length of d)*



(c) Eyelid distance of closed eye *(length of c ÷ length of d)*

Fig. 2. Extracting and normalising eye movement features: *After locating specific points of the eyes using an AAM, horizontal and vertical iris movements are extracted (a) and the eyelid distance is calculated as illustrated in (b) & (c).*

details, see [38]). Moreover, a total of 128 statistical features ("functionals") were extracted:

- Maximum, minimum, mean, variance, and standard deviation for all 18 low-level features mentioned earlier (5×18 features)
- Maximum, minimum, and average of: duration of looking left, right, up and down, as well as blink duration for each eye (3×2 eyes $\times$ 5 features)
- Closed eye duration rate, and closed eye to open eye duration rate for both eyes (2 eyes $\times$ 2 features)
- Blinking rate for both eyes (2 eyes $\times$ 1 feature)
- Gaze aversion rate for both eyes (rate of duration of non-frontal gazing) (2 eyes $\times$ 1 feature).

3.2.3 Head Pose and Movement Features

Simple behaviours such as head movements could provide useful cues about the mood, emotions, personality, or cognitive processing as found in [39], [40], [41], [42]. Previously [43], we found slower and less frequent head movements, increased eye contact avoidance and less social engagement with the clinical examiner, likely to also show in other social interactions.

To extract head pose and movement behaviour, the face had to be detected and tracked before a 3 degrees of freedom (DOF) head pose could be calculated (yaw, roll and pitch). We trained and built a subject-specific face active appearance model, where 30 images per subject were selected for manual annotation, then used for the face model. A 3D face model was projected onto our 2D face AAM to estimate the 3-DOF head pose (for more details refer to [43]).

These 3-DOF pose features, as well as their velocity and acceleration, were extracted to give a total of nine low-level features per frame. Over the duration of each subject's interview, a total of 185 statistical features were extracted:

- Maximum, minimum, range, mean, variance, and standard deviation for all nine low-level features mentioned earlier. (6×9 features)
- Maximum, minimum, range and average duration of: head direction left, right, up and down, tilting clockwise and anticlockwise. (4×6 features)
- Head direction duration rate, and rate of different head directions for non-frontal head direction for all directions mentioned above. (2×6 features)
- Change of head direction rate for all directions mentioned above. (1×6 features)
- Total number of changes of head direction for yaw, roll, pitch, and all directions. (1×4 features)
- Maximum, minimum, range, mean, variance, duration, and rate for slow, fast, steady, and continuous movement of yaw, roll, pitch. ($7 \times$ 3-DOF $\times 4$ features)
- Average duration of head aversion (average duration of non-frontal head direction) (1 feature).

The above eye and head duration features were detected when the feature in question is higher or lower than the average of the feature in question plus or minus the standard deviation of that feature for each subject's interview. For example, blink is detected as follows:

$$Blink = \begin{cases} 1 & : x < \mu - \sigma \\ 0 & : x > \mu - \sigma. \end{cases}$$

Where x is the normalised distance between the eyelids (length of c divided by the length of d (see Fig. 2)) and μ and σ are its mean and standard deviation respectively.

3.3 Analysis and Evaluation

3.3.1 Classification

For classification results reported in this work, we used SVM classifiers, which are discriminative methods that learn boundaries between classes. SVM has been widely used in emotion classification tasks [44], and often considered state-of-the-art, since it provides good generalisation properties [45]. To increase the accuracy of the results of SVMs, the cost and gamma parameters were optimised. We used LibSVM [46] to this end, with a wide range grid search for the best parameters with a radial basis function (RBF) kernel. To optimise the cost and gamma parameters, double-cross validation was used. That is, for each training turn in leave-one-out cross-validation, an inner 10-fold cross-validation was used. The final selected parameters are the ones that generalised to all training observations with all turns of the leave-one-out cross-validation. In other words, the common parameters that give the highest training average recall of all training sets in the cross-validation models were selected. This overall optimisation was performed to overcome the overfitting to the training sets, and therefore is able to generalise to the testing sets. The use of the average recall in optimising the parameters is to give a balance of recall measure for each class. That is, parameter selection finds as similar as possible correct classification results in both classes.

The classification was performed in a binary (i.e., depressed/non-depressed) subject-independent scenario. Given the relatively modest number of (depressed and control) subjects, which is a common problem in similar studies, a separate development set could not be used. To mitigate this limitation, a leave-one-subject-out cross-validation was used for classification, feature selection, and fusion without any overlap between training and testing data. Therefore, the results might be somewhat optimistic.

We measure the performance of the system in terms of average recall (AR), to be consistent with the parameter optimisation procedure explained above. Moreover, AR considers the correct recognition in both groups (depressed/non-depressed). The AR is also called "balanced accuracy", and is calculated as the mean of sensitivity and specificity. Since the SVM parameters are optimised using AR, the final classification results are balanced between the two binary classes. Therefore, measures from the confusion matrix (e.g., accuracy, F1 score) are approximately equal ($/pm 0.001$).

3.3.2 Statistical Analysis

In order to understand and characterise the behavioural patterns, the extracted statistical functional features from depressed and control groups were evaluated and compared statistically for significance. As our analysis was performed in a binary manner (depressed/non-depressed), a two-sample two-tailed T-test was used for statistical analysis purposes. The two-tailed T-tests assume unequal variances with significance $p = 0.05$. The sign of the T-test was also calculated to identify the direction of the effect.

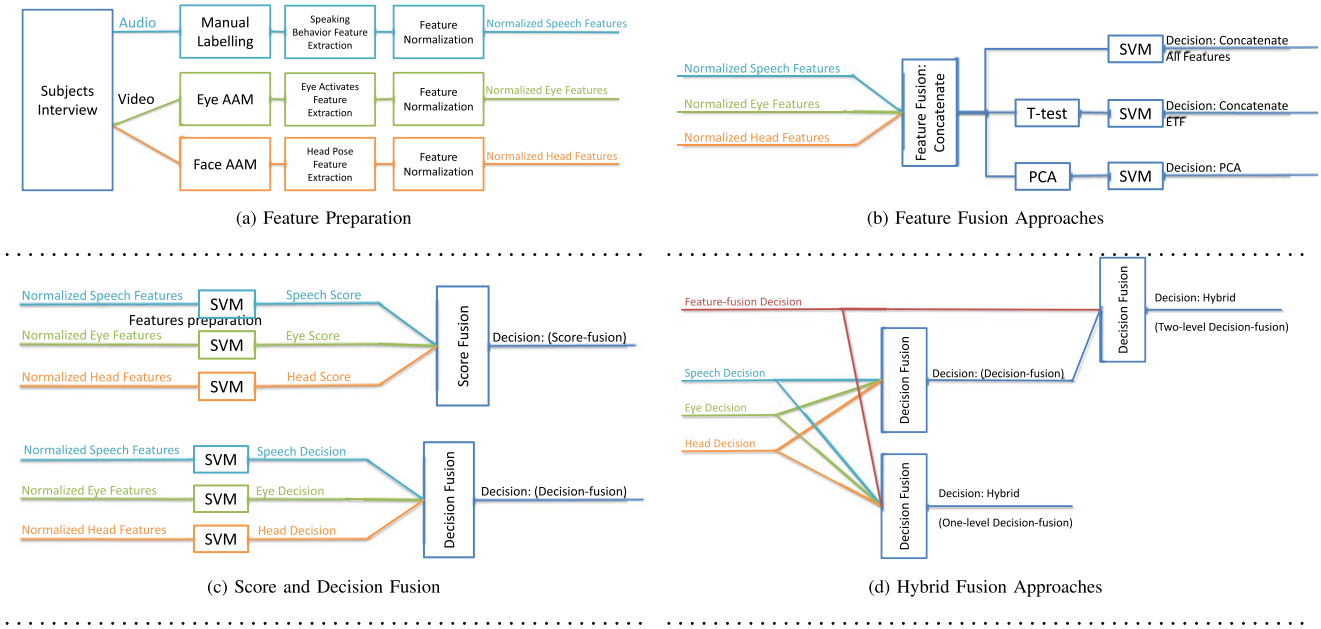


Fig. 3. Summary of the investigated system configurations: (a) feature preparation steps for audio and video data. Using the normalised features: (b) different feature fusion approaches were investigated, (c) score and decision fusions were performed after SVM classification, (d) hybrid fusion combined the decisions from the feature fusion with the decisions from the individual modalities in two approaches (one-level and two-level decision fusions).

Worth noting is that the threshold of $p < 0.05$ is uncorrected for the multiple comparison problem, which is due to several reasons. First, the tests are not conducted on the same raw data (i.e., different modalities). Second, correction for multiple comparison is not needed for feature analyses rather for feature interpretation [47], [48]. Since the tests are performed for feature selection for the classification purposes, a correction is not needed.

3.3.3 Feature Selection and Dimensionality Reduction

In order to maximise the recognition rate measured by AR, we experimentally compared two feature selection methods on the extracted statistical features.

First, the statistical test T-test mentioned above was used to rank the power of each attribute individually. Therefore, we selected features that exceeded the T-statistics (ETF) for being significantly different in the two groups (depressed/non-depressed) (i.e., p-value < 0.05). We initially performed this task using three approaches: (1) select ETF based on all subjects' data before classification, (2) select ETF of training set in each cross-validation turn, then apply them on the testing set, where the selected ETF will be variable in every turn, and (3) select only the mutual ETF of each training set in each cross-validation turn. Preliminary experiments showed that overfitting is a potential problem in all three methods. This warrants further investigation in the future but is beyond the scope of this study. Nevertheless, the classification results were not significantly different between the three methods. Therefore, in subsequent analysis, we used the *second* approach of ETF to reduce overfitting, to maintain the subject-independent approach, and to ensure a fair comparison for each cross-validation turn.

The second method investigated was a feature transformation method using PCA. PCA was performed on the extracted features of individual modalities and then on the combined features of the fused modalities for improvement comparison (i.e., to compare improvement and contribution from each modality with fused modalities). The use of the PCA for this study was not only for dimensionality reduction of concatenated features, but also to use the most promising principal components (PCs) that have the largest variance (98 percent of variances) to investigate their effect on recognition rate.

3.3.4 Normalisation

When modelling inputs with different scales, normalisation is recommended [49], which is the case in our study. In this work, min-max normalisation (scaling between 0 and 1) was performed, which is a linear transformation. It has the advantage of preserving exactly all relationships in the data by only changing the scale. However, it does not reduce the effects of outliers.[3] All classifications, statistical tests as well as PCA were performed on the normalised features.

3.4 Fusion

Multimodal fusion of different modalities can improve the classification performance, as it provides more useful information compared to using only a single modality. Fusion can be performed as pre-matching (early) fusion and post-matching (late) fusion. As one of the main objectives of this study is to investigate the best fusion approach for the classification of depression, three levels of fusion-feature, score, and decision—were investigated. Fig. 3 summarises the method of investigation used in this study.

- *Feature fusion:* Several methods were implemented in this study: (1) simply concatenating all extracted features from individual modalities, (2) concatenating exceeded T-test features (ETF) features from individual modalities, and (3) performing a PCA over the

3. Z-score normalisation was also performed, with the results being similar to min-max normalisation.

TABLE 2
Average Recall (AR) Classification Results for Individual Modalities

Row	Individual Modalities	#	Speech	#	Eye	#	Head	Average
1	All features	88	81.7	126	68.3	184	66.7	72.2
2	ETF	42-49	83.3	16-23	73.3	8-15	63.3	73.3
3	PCA	27	66.7	36	76.8	5	70.1	71.2

#: *Number of features in a super-vector.*

concatenated features (see Fig. 3b). PCA was applied on the features extracted from individual modalities as well as the concatenated features in the fused modality. 98 percent of the PCA variance was kept (empirically chosen), to compare improvement and contribution from each modality and the usability of dimensionality reduction.

- *Score fusion:* Several score techniques were implemented using the distance from the SVM hyperplane as scores (see Fig. 3c). Simple score fusion techniques (i.e., sum-rule, and product-rule) as well as secondary classifier (i.e., SVM) were applied.
- *Decision fusion:* Similarly to the score fusion, decision fusion fuses decisions (labels) of individual modalities (see Fig. 3c). For decision fusion, majority voting, and logical OR, as well as a secondary SVM classifier were also used for comparison.
- *Hybrid fusion:* This method can employ the advantages of both early and late fusion strategies. Therefore, hybrid fusion is investigated in this work to explore its results improvement and suitability for our multimodal depression detection goal. Majority voting and secondary SVM classifier at both one level and two levels decision fusion were used for hybrid fusion investigation purposes (see Fig. 3d). One-level decision fusion treats the feature fusion as an independent modality, where it fuses decision from feature fused modalities with decisions from individual modalities using one stage of decision fusion. Two-level decision fusion uses two stages of decision fusion such that: (1) a first stage decision fusion fuses the decision from individual modalities, then (2) a second stage decision fusion to fuse the first stage decision with the decision from the feature fused modalities.

Since a larger database with the desired characteristics (e.g., clinically validated, gender balanced, etc.) was not available to us for this task, weighted and complex fusion approaches could not be implemented in this study.

4 Results

Extracted statistical features from speaking behaviour, eye activities, and head pose were analysed individually and following fusion. These features were evaluated statistically to select the most promising ones to be used in the binary classification experiments.

Tables 2 and 3 show the classification results from individual modalities as well as when fused using different fusion methods, respectively. For each individual modality, we compared classification using (1) all extracted features, then with feature dimensionality reduction using: (2) only features that exceeded the T-statistics (ETF) (i.e., $p < 0.05$), and (3) feature transformation using PCA. All results presented are Average Recall (AR) rates (see Section 3.3.1). In general, using ETF performed best compared to the other methods, where individual modality classification results were already considerably higher than chance level (83 percent for speech, 73 percent for eye, and 63 percent for head modalities). This might be due to the statistical nature of the extracted features.

Classification rates from the speech modality using all features, and ETF have similar results, even with the reduction in feature vector size. For eye and head individual modalities, ETF reduced the feature vector enormously. Yet for the eye modality, the reduction had an influence in the

TABLE 3
Average Recall (AR) Classification Results for Fused Modalities

Row	Fused Modalities		Speech + Eye	Speech + Head	Eye + Head	All modalities	Method Average
4	Feature Fusion	Concatenate All features	214 81.7	272 ***88.3***	310 70.4	398 80.0	80.1
5		Concatenate ETF	59-70 80.0	51-61 83.3	24-37 70.0	69-83 80.0	78.3
6		98% of var. of PCA over concatenated features	47 81.7	46-47 85.0	49-50 65.4	51 78.3	77.6
	The rest of fusion methods are performed over the results obtained by ETF						
7	Score Fusion	Sum-rule	81.7	63.3	63.3	63.3	67.9
8		Product-rule	46.7	60.0	50.0	53.3	52.5
9		SVM	**85.0**	**86.7**	**78.3**	**86.7**	84.2
10	Decision Fusion	Majority	76.7	78.3	68.3	83.3	76.7
11		OR	80.0	68.3	68.3	66.6	70.8
12		SVM	83.3	83.3	73.3	**85.0**	81.2
13	Hybrid: Two-level	Majority	76.7	78.3	70.0	81.7	77.8
14		SVM	83.3	85.0	**74.7**	**85.0**	82.0
15	Hybrid: One-level	Majority	80.0	85.0	73.3	81.7	80.0
16		SVM	83.3	***88.3***	73.3	83.3	82.1
	Fused Modalities Average		78.5	79.5	69.1	77.6	-

#: *Number of features in a super-vector.* Bold: *highest AR for fusion compared to best result from individual modalities.* Italic: *highest AR overall. Method average is the row average to show the average of each fusion method. Fused modalities average is the column average to show the average of each modality combination.*

improvement of classification results compared to using all features. Typically, feature selection reduces complexity and dimensionality, which could improve the classification results slightly based on a comparative study [50]. Moreover, improvements in classification results when using feature selection depend on several factors including classifier, feature selection algorithm, and dataset [50]. Since these factors are constant for the three modalities, the considerably improved classification result from eye modalities suggests that irrelevant features could confuse and reduce the recognition rate of the classifier.

We also performed a PCA over the extracted features from individual modalities (see row #3). The results shows a considerably lower recognition rate for speech modality compared with the results from using all features and the statistically selected features (rows #1 and #2). For eye and head modalities, a slightly higher recognition rate was obtained from using PCA compared with the results from using all features and ETF (rows #1 and #2). These inconstancy between the results of investigated modalities suggest that even the top 98 percent of PCA variances do not have similar discriminative power as in the features selected by ETF or all features in these cases. Thus, all, ETF, and PCA features combination effectiveness in depression recognition rate will be invigilated further in the fused modalities.

Fusion approaches differ in when and how to fuse the modalities in question (see Section 3.4). While early fusion could be executed as sensor fusion or feature fusion, late fusion is executed either as decision or score fusion. As the name implies, hybrid fusion is a mixture of both early and late fusion. Early fusion, late fusion or a hybrid of both, with different methods of each are investigated. The result of each method is shown in Table 3. Moreover, we inspected all possible combinations of the three modalities to observe the contribution of each modality in the fusion process. We anticipated that fusing modalities will improve the performance compared to its individual modalities. Moreover, we hypothesised that when using all modalities, modality fusion would not only improve the results from the individual modalities, but also increase the confidence level of the final decision.

Regardless of the method, fusing these modalities results in either higher or at least not catastrophic compared to individual modality results (with the exception of product-rule). In general, score fusion using secondary classifier yielded the highest and the most robust classification rate (84 percent average accuracy). However, since a secondary classifier might risk overfitting, a larger database to validate this results is needed.

Early fusion, in particular feature fusion, was performed (see Fig. 3b for a visual illustration) by (1) concatenating all extracted features from individual modalities (see row #4), (2) concatenating ETF features from individual modalities (row #5), and (3) performing a PCA over the concatenated features from fused modalities (row #6). With all but one modality combinations, classification results of fusing all extracted features (row #4) have a slight improvement compared with the results for individual modalities (row #1). One exception is when fusing all modalities a slight reduction in recognition rate occur. On the other hand, concatenating ETF decreases the results slightly, with one exception of remarkable improvement when fusing eye and head modalities (row #5).

Comparing the three feature selection methods in individual and combination of modalities, the average of the recognition result are similar, with ETF feature selection yielded the highest results. Since the extracted features are statistical to overcome normalisation issues, we believe that statistical feature selection methods best fit the purpose.

We believe that as we are extracting statistical features from each modality, the feature fusion technique drawbacks are reduced, such as synchronisation, frame rate and dimensionality differences. As expected, feature normalisation made features from different modalities compatible to be combined, which reduces classifier bias towards some features rather than others. Moreover, feature selection techniques were able to pick the most promising features for the classification task.

To reduce the complexity for the following fusion approaches, we selected the results performed by ETF for individual modalities (see row #2), as well as ETF results obtained by fusing all modalities (see row #5) for hybrid fusion. The choice of ETF is due to its stability and improvement for both individual and fused modalities (see rows #2 & #5).

With late fusion, we explored score and decision fusion using several approaches (see rows #7-12). Score fusion is explored using sum-rule, product-rule, max-rule and secondary SVM methods, where the scores are the distances from the SVM hyperplane. On the other hand, decisions (labels) out of individual modality classifications were also fused using majority voting, logic AND, logic OR, as well as a secondary SVM. As expected, max-rule performance was exactly the same as logic OR performance, therefore only logic OR is shown in the results table. The same applies to product-rule and logic AND, where only product-rule is shown in the results table. This similarity might imply that using the distance from the SVM hyperplane as scores is similar to using their decision label. Therefore, using classification scores might be applicable when fusing different types of classifiers. Majority voting with an even number of votes performs similarly to logical AND when number of votes of each class are equal.

In general, traditional methods of score and decision fusion results did not improve over individual results, yet at least the results were not catastrophic (i.e., not worse than the lowest individual modality results), except for Product-rule. Worth noting is that signs (positive and negative scores) are used to identify the class that the subject is classified as belonging to along with the score, which is the distances from the SVM hyper plane. That is, a positive score is given to the depressed class and a negative score is given to the control class. Therefore, mathematical operations that rely on the sign of the scores (i.e., max-rule and product-rule) of individual modalities affect the mathematical sign of the fused modality. For example, if a control subject is misclassified as depressed even for only one modality (a positive sign), the the multiplication operation in the product-rule fusion classification will result in classifying that subject as depressed regardless of the classification of the other modalities. Therefore, the catastrophic results obtained by the product-rule score fusion methods might be due to the effect of the mathematical operations on the acquired signs of the scores.

Nevertheless, having no improvement in AR is not an indication of low performance on its own, as it might

increase the decision confidence level. On the other hand, using a secondary SVM has improved the classification results in both score and decision fusion. Here, the predicted classification scores or the labels are used as feature vector, also in a leave-one-out cross validation.

Hybrid fusion was investigated for the benefits of utilising both feature fusion and decision fusion. In this study, hybrid fusion was also examined in two ways in order to inspect its effectiveness for our multimodal depression detection. First, using two levels of decision fusion (rows #13-14). Second, using one-level decision fusion (rows #15-16). We used majority voting and a secondary SVM as decision fusion for both one-level and two-levels hybrid fusion methods. Majority voting was chosen over other logic functions (e.g., AND), as it seemed more reliable especially with more votes, while for the two-level hybrid fusion, where there are two votes, it acted as logic AND.

For hybrid majority voting, even though none improved over individual modalities, one-level (see row #15) performed better than two-level (see row #13) hybrid fusion due to having more votes to decide upon. Knowing that majority voting performs similarly to logical AND with an even number of votes, one- and two-level hybrid fusion result in similar decisions when fusing all three modalities. Even though they perform similarly, we believe that the majority voting of one-level hybrid fusion is more reliable and more robust to overfitting than the two-level one. This is because with one-level hybrid fusion, we end up with four votes, while with two-levels we end up with two votes.

Acknowledging the risk of overfitting, we performed a hybrid fusion with secondary SVM on decisions from individual modalities and fused modalities in a one- (see row #16) and two- (see row #14) levels decision fusion. Most of the cases either slightly improved the results or at least matched the higher individual modality result. As a secondary SVM might risk overfitting, it would likely need a larger database for this approach to be validated.

The last column of Table 3 shows the method average. As can be seen, the highest classification average of the fusion methods is achieved when using a secondary classifier with the classification scores. However, since using a secondary classifier might risk overfitting, the result should be validated using a dataset with a large number of samples.

The last row of Table 3 shows the fused modalities average, where the average classification results of each column is calculated to show the average of each modality combination. As it can be seen, none of the average classification results of the modalities combination outperformed the highest classification results of their individual modality classification results (row #2 Table 2). Yet, none of them was worse than the lowest classification results of their individual modality classification results. Even though the modalities combination classification average did not improve compared to the individual modalities, it increases the confidence level of the final decision.

As mentioned in the background section, comparison of our current fusion results with previous multimodal depression detection studies is difficult due to differences in recording environment (e.g., equipment, paradigm, etc.), and methodology (e.g., extracted features, classification label, fusion methods, etc.). Regardless of methodology differences, the general improvement in the results when fusion techniques are used in this study is in line with the improvement in results of the previous studies [19], [20], [21], [22], [23]. Despite the difference in type of extracted features in our previous study [23], the results from different fusion techniques show similarity with this current study. Similarly to current results, in [23] a secondary classifier produced higher results than other fusion techniques of the same fusion level. Unlike [23] where using PCA in feature fusion performed the best compared to other feature fusion techniques, it performed the worst in this study, which might be due to the differences in the type of investigated features (low-level features versus functional features).

TABLE 4
Number of Misclassified Subjects in Each Modality

Modality		Speech	Eye	Head	Concatenated ETF
Depressed	Males	3	5	5	5
	Females	2	3	2	1
Control	Males	3	5	9	5
	Females	2	3	6	1

5 Error Analysis

For a better understanding of the misclassifications in each modality, we analysed the errors based on subjects' metadata. Table 4 shows the number of subjects that have been misclassified in each modality (speech, eye, head) as well as feature fusion when using ETF.

As can be seen, for each modality, the chance of misclassifying males is higher than for females for both depressed and control groups (except for the head modality). This result is consistent with previous findings of gender differences [51] that depression in women may be more likely to be detected than depression in men. We speculate that this might be related to the theory that women are more likely to amplify their mood [51]. The same study suggested that men are more likely to engage in distracting behaviours that dampen their mood when depressed. However, that does not explain the misclassifications of male control subjects.

For the speech modality, as our features are behavioural in nature (e.g., response time), signal quality and gender-dependent feature issues are eliminated. The number of misclassified depressed and controls from both male and female are equivalent. That might be due to the optimisation of SVM parameters, where the AR results is higher with balanced classification rate from the two classes than unbalanced classification. Nevertheless, female subject misclassifications is less than the misclassification of male subjects of both classes.

With eye and head modalities, we explored the effect of video quality and whether the subject was wearing glasses. Regarding video quality, only three videos had slightly blurred images (all from the control subset). All three have misclassifications from the head modality and only one has misclassifications from the eye modality. However, that does not explain the misclassifications from normal quality videos. Besides, as the eye-AAM and face-AAM were annotated and trained in a subject-dependent manner, it is also dependent on the recording conditions. Therefore, we believe that the misclassifications were not based on the quality of the video or the method used.

We also looked at errors based on age, diagnosis, depression score, medications (current and history), family history, smoking and alcohol consumption; none of which had an effect on the classification errors. Moreover, Australia is a multicultural country, therefore, even with selecting native Australian English speakers, three subjects have an Asian heritage appearance (all control subjects: one older male and two young females). While none of the Asian young females were misclassified in most modalities (only one of the females was misclassified using the head modality), the Asian male was misclassified in each modality and combined modalities. As there is not enough data to draw a conclusion, future work could investigate the influence of cultural backgrounds.

Therefore, we believe that, as all extracted features were behavioural in nature, subjects of certain personalities and backgrounds might act and behave differently regardless of their mental health. For example, depressed patients who have more head movement as they speak are misclassified as control subjects and vice versa. The same applies for the eye and speech modalities. As the current data collection did not include personality assessment, we could not derive a solid conclusion, which is also being considered for our ongoing data collection. Nevertheless, the current error analysis is rudimentary, where a formal and statistical based analysis is needed to validate these results. Since the scope and focus of this paper is on classification, future work should advance such error analysis, as performed in [52].

6 Conclusion

Intending to ultimately develop an objective multimodal system that supports therapists during the diagnosis and monitoring of clinical depression, we investigated verbal and nonverbal statistical patterns of depression individually and when fused. To develop a classification system-oriented approach, this paper conducted feature selection, classification and fusion-based experiments to conclude which combinations of behaviour (verbal and nonverbal) can best discriminate between depression and non-depression. We analysed the statistical significance of each feature of depression behaviour from these modalities to select the most relevant features for classification.

We examined the performance of binary classification using these modalities individually and when fused in different combinations. An SVM classifier was used for classification using several feature selection methods and several fusion approaches. Given the statistical nature of the extracted features, using T-test as feature selection method performed best compared to the other methods, where individual modality classification results were already considerably higher than chance level (83 percent for speech, 73 percent for eye, and 63 percent for head modalities).

When fusing these modalities using different fusion methods, the results were either higher or at least not catastrophic compared to individual modality results (with the exception of product-rule). Among the different fusion approaches used here, the highest and the most robust fusion method was score fusion using secondary classifier, giving up to 84 percent average accuracy. As a secondary classifier of score, decision, and hybrid fusion might risk overfitting, it would likely need a larger database for this approach to be convincing and valid.

Finally, we have analysed the classification errors for a better understanding of our method for detecting depression. In line with the literature, depression in women is more likely to be correctly classified than in men from either group (depressed and control). We eliminated several technical issues that might have an effect on the classification, including audio and video quality, gender-dependent features, and recording conditions. Moreover, none of the subjects' meta-data had an effect on the classification. It should also be noted that depression diagnosis is also based on clinical history, not just mental state examination, which was the focus of this present study.

7 Limitations and Future Work

Even though it is a common problem in similar studies, a known limitation is the relatively modest number of (depressed and control) subjects because of the selection criteria imposed in this paper. A large-scale study using clinically validated depression diagnosis is preferable, however, to the best of our knowledge, is not available. Crowd-sourcing is one means of sourcing very large amounts of data that is growing in popularity, but crowd-sourcing such depression datasets not only could lack the clinical assessment, which we believe is crucial, but also might lack variety of depression severity scores. As the Black Dog Institute data collection is ongoing, we anticipate reporting on a larger dataset in the future. Moreover, future data collection will aim to match and model the ethnicity and culture to investigate their influence on the expression of emotion and on automated depression detection.

Moreover, to get as accurate features as possible, we have used manual annotation for speech, and subject-specific AAM for eye and head (automatic feature extraction was not the focus of this study). Speech annotation and speaker separation could be attempted automatically using advanced speaker diarisation techniques. Regarding eye activity features, automated algorithms that measure blink, eyelids and iris movements could be utilised for this task. For head pose and movement, a general face tracker could be effective in extracting head pose features. Therefore, having a fully automated system to extract and analyse the proposed features is feasible for the task of detecting depression but was not the focus of this study.

In this work, depression was investigated in a binary classification manner (i.e., severe depressed versus healthy controls). However, having a regression classification problem to detect depression severity could be a next step for an advanced depression diagnosis system. Such a regression problem needs a large dataset with a variety of depression severity scores, as mentioned above, noting the difficulty of obtaining an agreed severity score from clinical assessment. Furthermore, fusing prediction scores of a regression classification requires different fusion methods than the ones used for fusion of classifiers. While the current study focused on between-subjects design, a within-subjects design would assist in longitudinal monitoring of depression. Future automated depression monitoring studies could consider this promising analysis.

Further fusion approaches in recent years have been presented by [53], [54], [55]. It would be worthwhile to explore

their usage in the context of multimodal depression detection in the future, as it is beyond the scope of this paper. Moreover, other features such as vocal prosody (e.g., energy, pitch) and facial expressions could be extracted and fused with the current investigated approach. This study focused on extracting, analysing, and selecting behavioural patterns of subjects' responses, where speech behaviour, eye activity, and head movement were investigated. Vocal prosody, facial expressions, and body posture modalities were not included in this paper, but are acknowledged as potential sources of information. However, future work will investigate such modalities, seeking more accurate and confident diagnoses of depression. Moreover, the findings of the current study will be validated (using the same protocol) for generalisation across cultures (American, German, Saudi) and languages (American-English, German, Arabic) using different datasets.

Acknowledgments

This research was funded in part by the Australian Research Council (ARC) Discovery Project grant DP130101094.

References

[1] C. Mathers, J. Boerma, and D. Fat, *The Global Burden of Disease: 2004 Update*. Geneva, Switzerland: WHO, 2008.

[2] S. B. Guze and E. Robins, "Suicide and primary affective disorders," *British J. Psychiatry*, vol. 117, no. 539, pp. 437–438, Oct. 1970.

[3] L. G. Kiloh, G. Andrews, and M. Neilson, "The long-term outcome of depressive illness," *British J. Psychiatry*, vol. 153, no. 6, pp. 752–757, 1988.

[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] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed. Arlington, TX, USA: American Psychiatric Association, 2000.

[6] N. Sebe, I. Cohen, and T. S. Huang, "Multimodal emotion recognition," *Handbook Pattern Recognit. Comput. Vis.*, vol. 4, pp. 387–419, 2005.

[7] S. D'Mello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Model. User-Adapted Interaction*, vol. 20, no. 2, pp. 147–187, 2010.

[8] M. Valstar, et al., "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, 2013, pp. 3–10.

[9] S. D'Mello and J. Kory, "Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proc. 14th ACM Int. Conf. Multimodal Interaction*, 2012, pp. 31–38.

[10] A. Rattani, D. R. Kisku, M. Bicego, and M. Tistarelli, "Feature level fusion of face and fingerprint biometrics," in *Proc. IEEE Conf. Biometrics: Theory, Appl. Syst.*, 2007, pp. 1–6.

[11] J. Crowley, "Principles and techniques for sensor data fusion," in *Multisensor Fusion for Computer Vision*, J. Aggarwal, Ed. Berlin, Germany: Springer, 1993, vol. 99, pp. 15–36.

[12] J. Movellan and P. Mineiro, "Robust sensor fusion: Analysis and application to audio visual speech recognition," *Mach. Learn.*, vol. 32, no. 2, pp. 85–100, 1998.

[13] H. B. Mitchell, *Multi-Sensor Data Fusion: An Introduction*. Berlin, Germany: Springer, 2007.

[14] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[15] K. Pearson, "On lines and planes of closest fit to systems of points in space," *London Edinburgh Dublin Philosophical Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.

[16] S. Tulyakov, S. Jaeger, V. Govindaraju, and D. Doermann, "Review of classifier combination methods," in *Machine Learning in Document Analysis and Recognition*. Berlin, Germany: Springer, 2008, pp. 361–386.

[17] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010.

[18] J. F. Cohn, et al., "Detecting depression from facial actions and vocal prosody," in *Proc. 3rd Int. Conf. Affective Comput. Intell. Interaction Workshops*, 2009, pp. 1–7.

[19] S. Scherer, G. Stratou, and L.-P. Morency, "Audiovisual behavior descriptors for depression assessment," in *Proc. 15th ACM Int. Conf. Multimodal Interaction*, 2013, pp. 135–140.

[20] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: A multimodal approach," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, 2013, pp. 11–20.

[21] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 65–72.

[22] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, 2013, pp. 21–30.

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

[24] D. V. Sheehan, et al., *The Mini-International Neuropsychiatric Interview (M.I.N.I.): The Development and Validation of a Structured Diagnostic Psychiatric Interview for DSM-IV and ICD-10*, vol. 59, no. suppl. 20. Memphis, TN, USA: Physicians Postgraduate Press, 1998, pp. 22–33.

[25] D. Wechsler, *Wechsler Test of Adult Reading: WTAR*. San Antonio, TX, USA: Psychological Corporation, 2001.

[26] A. J. Rush, et al., "The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression," *Biol. Psychiatry*, vol. 54, no. 5, pp. 573–583, 2003.

[27] H. Ellgring and K. R. Scherer, "Vocal indicators of mood change in depression," *J. Nonverbal Behavior*, vol. 20, no. 2, pp. 83–110, 1996.

[28] C. Sobin and H. A. Sackeim, "Psychomotor symptoms of depression," *Amer. J. Psychiatry*, vol. 154, no. 1, pp. 4–17, 1997.

[29] A. J. Zlochower and J. F. Cohn, "Vocal timing in face-to-face interaction of clinically depressed and nondepressed mothers and their 4-month-old infants," *Infant Behavior Develop.*, vol. 19, no. 3, pp. 371–374, 1996.

[30] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "From joyous to clinically depressed: Mood detection using spontaneous speech," in *Proc. FLAIRS-25*, 2012, pp. 141–146.

[31] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting depression: A comparison between spontaneous and read speech," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2013, pp. 7547–7551.

[32] S. Alghowinem, et al., "A comparative study of different classifiers for detecting depression from spontaneous speech," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2013, pp. 8022–8026.

[33] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Characterising depressed speech for classification," in *Proc. Interspeech*, 2013, pp. 2534–2538.

[34] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Res. Methods*, vol. 41, no. 2, pp. 385–390, 2009.

[35] F. Goldman-Eisler, *Psycholinguistics: Experiments in Spontaneous Speech*. New York, NY, USA: Academic, 1968.

[36] H. Ellgring, *Non-Verbal Communication in Depression*. Cambridge, U.K.: Cambridge Univ. Press, 1989.

[37] J. H. Mackintosh, R. Kumar, and T. Kitamura, "Blink rate in psychiatric illness," *British J. Psychiatry*, vol. 143, no. 1, pp. 55–57, 1983.

[38] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Eye movement analysis for depression detection," in *Proc. IEEE Int. Conf. Image Process.*, 2013, pp. 4220–4224.

[39] D. Heylen, "Head gestures, gaze and the principles of conversational structure," *Int. J. Humanoid Robot.*, vol. 3, no. 3, pp. 241–267, 2006.

[40] J. Pedersen, J. T. M. Schelde, E. Hannibal, K. Behnke, B. M. Nielsen, and M. Hertz, "An ethological description of depression," *Acta Psychiatrica Scandinavica*, vol. 78, no. 3, pp. 320–330, 1988.

[41] L. Fossi, C. Faravelli, and M. Paoli, "The ethological approach to the assessment of depressive disorders," *J. Nervous Mental Disease*, vol. 172, no. 6, pp. 332–341, 1984.
[42] W. W. Hale III, J. H. Jansen, A. L. Bouhuys, J. A. Jenner, and R. H. van den Hoofdakker, "Non-verbal behavioral interactions of depressed patients with partners and strangers: The role of behavioral social support and involvement in depression persistence," *J. Affective Disorders*, vol. 44, no. 2/3, pp. 111–122, Jul. 1997.
[43] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Head pose and movement analysis as an indicator of depression," in *Proc. Humaine Assoc. Conf. Affective Comput. Intell. Interaction*, 2013, pp. 283–288.
[44] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[45] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, pp. 1062–1087, Feb. 2011.
[46] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, 2011.
[47] D. J. Saville, "Multiple comparison procedures: The practical solution," *Amer. Statistician*, vol. 44, no. 2, pp. 174–180, 1990. [Online]. Available: http://www.jstor.org/stable/2684163
[48] K. J. Rothman, "No adjustments are needed for multiple comparisons," *Epidemiology*, vol. 1, no. 1, pp. 43–46, 1990. [Online]. Available: http://www.jstor.org/stable/20065622
[49] T. Jayalakshmi and A. Santhakumaran, "Statistical normalization and back propagation for classification," *Int. J. Comput. Theory Eng.*, vol. 3, no. 1, pp. 1793–8201, 2011.
[50] E. M. Karabulut, S. A. Zel, and T. Briki, "A comparative study on the effect of feature selection on classification accuracy," *Procedia Technology*, vol. 1, pp. 323–327, 2012.
[51] S. Nolen-Hoeksema, "Sex differences in unipolar depression: Evidence and theory," *Psychol*, vol. 101, pp. 259–282, 1987.
[52] J. M. Girard, J. F. Cohn, L. A. Jeni, M. A. Sayette, and F. De la Torre, "Spontaneous facial expression in unscripted social interactions can be measured automatically," *Behavior Res. Methods*, vol. 47, no. 4, pp. 1136–1147, 2015.
[53] Y. Song, L. Morency, and R. Davis, "Multi-view latent variable discriminative models for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2120–2127.
[54] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Continuous conditional neural fields for structured regression," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 593–608.
[55] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2015.

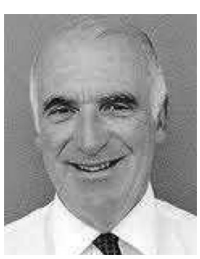
Sharifa Alghowinem received the BSc degree in computer applications from King Saud University, in 2004, the MSc degree in software engineering from the University of Canberra, in 2010, and the PhD degree from the Australian National University, Computer Science Research School, in 2015. Her research interests include speech processing, computer vision, affective computing, and machine learning. She worked as a lecturer with the University of Canberra, in 2011, and currently holds a research and teaching position with Prince Sultan University.

Roland Goecke received the master's degree in computer science from the University of Rostock, Germany, in 1998 and the PhD degree in computer science from the Australian National University, Canberra, Australia, in 2004. He is a professor of affective computing, head of the Vision and Sensing Group, and director of the Human-Centred Technology Research Centre, University of Canberra. Before joining the University of Canberra in 2008, he worked for Seeing Machines, National ICT Australia and the Fraunhofer Institute for Computer Graphics, Germany. His research interests include affective computing, pattern recognition, computer vision, human-computer interaction and multimodal signal processing. He is a member of the IEEE.

Michael Wagner received the MSc degree in physics (diplomphysiker) from the University of Munich, in 1973 and the PhD degree in computer science from the Australian National University, in 1979. He has held research and teaching positions with Technical University of Munich, National University of Singapore, University of Wollongong, University of New South Wales, Australian National University and University of Canberra as well as commercial research and development positions with Nixdorf AG, Germany and ANUTech, Australia. He is now Emeritus professor of computing with the University of Canberra and honorary professor with Australian National University and Technical University of Berlin.

Julien Epps received the BE and PhD degrees from the University of New South Wales, Australia, in 1997 and 2001, respectively. After being a postdoctoral fellow with the University of New South Wales, he worked on speech recognition and speech processing research first as a research engineer with Motorola Labs and then as a senior researcher with National ICT Australia. He joined the UNSW School of Electrical Engineering and Telecommunications as a senior lecturer, in 2007 and is currently an associate professor. He has also held visiting academic and research appointments with the University of Sydney and the A*STAR Institute for Infocomm Research (Singapore). His research interests include emotion and mental state recognition from speech and behavioural signals, and genomic signal processing. He is a member of the IEEE.

Matthew Hyett received the bachelor's of science and postgraduate diploma degrees in psychology from Macquarie University, in 2005 and 2007, respectively, and the PhD degree the School of Psychiatry, University of New South Wales (UNSW), in 2015. He worked as a research assistant in the School of Psychiatry, UNSW and Black Dog Institute between 2006 and 2012 (and in 2015/2016). Whilst there, he coordinated a large study into cognitive and neurobiological markers of melancholic and non-melancholic depression, before moving to QIMR Berghofer Medical Research Institute in 2012 (Systems Neuroscience Group). He moved to Perth in February 2016 and currently works as a research fellow in the School of Psychology and Speech Pathology, Curtin University. His research interests include span mood and anxiety disorder classification and treatment, cognitive neuroscience, and computational psychiatry.



Gordon Parker is a scientia professor of psychiatry with the University of New South Wales, and for 10 years held the position of inaugural executive director of the Black Dog Institute in Sydney, Australia. For nearly two decades he was head of the School of Psychiatry, UNSW and director of psychiatry, Prince of Wales and Prince Henry Hospitals. He is internationally recognised for his research into the causes and phenomenology of depressive and bipolar (mood) disorders. Such research has been pivotal in challenging existing diagnostic approaches, and has contributed to optimising treatments across differing conditions. He received a Citation Laureate award in the field of Psychology/Psychiatry, in 2004. In 2007, he was elected as a fellow of the Academy of the Social Sciences in Australia, and in 2010 received the Officer of Order of Australia Award for distinguished service to psychiatry.

Michael Breakspear received the medical training in 1994 and the PhD degree in 2003 and post-doctoral training in physics in 2003-2006 from the University of Sydney before joining the School of Psychiatry, UNSW. He moved to QIMR Berghofer in 2009 where he leads the Systems Neuroscience Group. He is a consultant psychiatrist and chair of research for Metro North Mental Health and a psychiatrist in the Forensic Mental Health Service.