Interpretation of Depression Detection Models via Feature Selection Methods

Sharifa Alghowinem, Tom Gedeon, *Senior Member, IEEE,* Roland Goecke, *Senior Member, IEEE,* Jeffrey F. Cohn, and Gordon Parker

Abstract—Given the prevalence of depression worldwide and its major impact on society, several studies employed artificial intelligence modelling to automatically detect and assess depression. However, interpretation of these models and cues are rarely discussed in detail in the AI community, but have received increased attention lately. In this study, we aim to analyse the commonly selected features using a proposed framework of several feature selection methods and their effect on the classification results, which will provide an interpretation of the depression detection model. The developed framework aggregates and selects the most promising features for modelling depression detection from 38 feature selection algorithms of different categories. Using three real-world depression datasets, 902 behavioural cues were extracted from speech behaviour, speech prosody, eye movement and head pose. To verify the generalisability of the proposed framework, we applied the entire process to depression datasets individually and when combined. The results from the proposed framework showed that speech behaviour features (e.g. pauses) are the most distinctive features of the depression detection model. From the speech prosody modality, the strongest feature groups were F0, HNR, formants, and MFCC, while for the eye activity modality they were left-right eye movement and gaze direction, and for the head modality it was yaw head movement. Modelling depression detection using the selected features (even though there are only 9 features) outperformed using all features in all the individual and combined datasets. Our feature selection framework did not only provide an interpretation of the model, but was also able to produce a higher accuracy of depression detection with a small number of features in varied datasets. This could help to reduce the processing time needed to extract features and creating the model.

Index Terms—depression detection, multimodal analysis, feature selection, datasets generalisation.

1 Introduction

According to the World Health Organisation (WHO), major depressive disorders are an increasing global issue that leads to devastating consequences [1]. A person living with depression suffers enormously and functions poorly in everyday life tasks. Depression is a major contributor to the overall global burden of disease and is the leading cause of disability worldwide. Depression is strongly linked with non-communicable disorders, such as diabetes and heart disease, increased risk of substance use disorders, and at its worst, it can lead to suicide. Even though treatments for depression are effective, only 10% of depressed patients receive such treatments, where one of the barriers to effective care is inaccurate diagnoses. Misdiagnosed and untreated depression does not only affect the sufferer at a personal level, but also affects the employer and the government at an economic level.

Given the prevalence of depression disorder worldwide and its major impact, several studies attempted to automatically detect and diagnose depression by employing artificial intelligence modelling. Modelling depression in these studies varied in terms of investigated modalities, extracted features, modelling algorithms, etc. Moreover, depression datasets collected by these studies are also different in purpose, procedure, and environment. The use of modelling algorithms with such diverse modalities and their high dimensional feature spaces, restricts the ability to interpret its results. Moreover, such differences prevent generalisation and drawing solid conclusions about the effectiveness of the modelling. Feature selection techniques aim at reducing the dimensionality of the feature space in order to increase the efficiency of the modelling. However, the majority of such techniques focus on reducing the number of features used to select the features with highest class discriminative power, without giving an insight into what features are being selected. Similarly, depression detection modelling studies that used feature selection methods did not provide the selected features by these techniques, which is important to interpret and generalise a model to other datasets.

To improve the interpretability and generalisability of depression analysis modelling, we investigate a novel and robust framework that aggregates the results from different feature selection methods. That is, we utilise several feature selection methods to interpret a depression detection model, and analyse their effect on the classification and generalisation results. We extract behavioural and functional features from speech, eye activities and head movement from three depression datasets. By having a closer look at the features being selected by the feature selection methods, we hypothesise that; (1) there are features that are commonly selected by feature selection methods that strongly distinguish depressed behaviour, (2) that these commonly selected features are robust to randomness and are consistently selected within different thresholds of the same feature selection method, and (3) finding these features can be helpful for generalisable

- *S. Alghowinem is with Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA, with Prince Sultan University, Riyadh, Saudi Arabia and with the Australian National University, Canberra, Australia E-mail: sharifah@mit.edu*
- *T. Gedeon is with the Australian National University, Canberra, Australia. E-mail: tom.gedeon@anu.edu.au*
- *R. Goecke is with the University of Canberra, Canberra, Australia E-mail: roland.goecke@ieee.org*
- *J.F. Cohn is with the University of Pittsburgh, Pittsburgh, PA, USA. E-mail: jeffcohn@pitt.edu*
- *G. Parker is with the University of New South Wales, Sydney, Australia. E-mail: g.parker@unsw.edu.au*

modelling of depression behaviour independent of the datasets and recording process.

To best of our knowledge, this paper is the first to explore an extensive array of feature selection methods to find the strongest features for a depression diagnosis. The novelty of this paper is as follows:

- We perform a comprehensive analysis of feature selection techniques in the context of depression behaviour from different family groups of feature selection techniques; traditional flat feature, dynamic (streaming) feature, and structural based feature selection techniques.
- We propose a framework to aggregate the results of the feature selection techniques to increase the robustness to randomness of the selected features, using different stability measures. We also propose a new feature selection stability measure using between thresholds stability.
- We generalise the framework by applying it on different depression datasets to select the more representative features. Then, we investigate the effectiveness of these features by creating depression severity models using different depression datasets.

This exploration facilitates the interpretation of depression behaviour since it is applied on a clinically annotated matched control-depressed database using multimodal behavioural analysis. Identifying the most meaningful depression behaviour patterns is significant since behaviour has been associated with depression-related symptoms in psychology literature. For this purpose, the main contribution of this paper is the extensively validated framework that not only reliably identified the characteristics of depression behaviour, but also provided an understanding of recognition model performance and its generalisability.

2 Related Work

Automatic modelling of depression aims at providing an objective measure for a depression diagnosis. This approach addresses the issue of misdiagnosing depression, which is considered one of the main barriers to depression treatments. These studies investigated several cues of depression analysis including; facial expression, speech characteristics, head and body movement, brain signals, linguistic choices for text and speech, etc. The extracted features from such studies differ between each other even when analysing the same modality, as well as machine learning techniques used in these studies varies. Intensive reviews of these studies are presented in [2] for visual features and [3] for speech features. Recent research on depression modelling share the same differences.

For example, with the emergence of deep learning techniques, [4] utilised deep convolutional neural networks (DCNN) to model depression diagnosis from facial expression. They used facial appearance and dynamic facial movement as images to fine-tune two separate pre-trained DCNNs, then fused their results in score-fusion level. The results of their networks architecture outperform those of other studies. However, due to the complexity of deep learning techniques, it is difficult to interpret the features that contributed to such improvement in the performance. Moreover, deep learning techniques need a huge number of labelled observations. Therefore, it is common to use pre-trained networks. In that study, the pre-trained sample was of a general face recognition dataset. This might have an effect on the diagnosis of depression.

Another study modelled forecasting depression mood based on self-reported history using a recurrent neural network (RNN) [5]. The study used long-term historical information of a user that includes; user reported mood, action, medications, sleeping patterns, etc. The subjects of the study report the information periodically each day over several months. The study could accurately forecast severe depression mood up to two weeks in advance. However, since the information is self-reported, there is no indication of how such a model would work as an automated system.

Another example of deep learning utilisation for depression detection is presented in [6], where audio and text features are extracted from virtual agent-human interaction interviews. A long-short-term memory (LSTM) neural network model is built using audio and text features, where it outperformed the baseline result. Although, it is difficult without further investigation to interpret the model to know exactly what features contributed to the high performance.

Using AVEC dataset, [7] investigated three deep learning techniques as a method of transferred learning in an effort to increase the diagnosis of depression severity based on visual cues. Even though their results show similarity to the state-of-the-art, it does not provide an interpretation of the model and its effectiveness.

Even without using deep learning, automatic feature extraction from video signal are not always interpretable. For example, in [8], dynamic facial feature descriptors are automatically extracted from the video recording of AVEC depression dataset. Local binary pattern is extracted from three orthogonal planes to capture microstructure of facial appearance, where the results histograms are concatenated. Fisher vector is then used to cluster the features, and a support vector regression is used for depression severity modelling. The results showed an improvement from the baseline and other studies. However, such approaches are difficult to interpret.

Feature selection methods have been utilised for depression modelling studies, with the goal of improving the accuracy of depression diagnosis. Both survey papers of [2] and [3] listed some studies that utilised feature selection methods. However, such studies do not report the selected feature set, which would improve the understanding of the generalisation of their findings, nor report stability measures and the procedure to increase it. Moreover, some of these studies used feature transformation methods, where the actual features that contribute to the modelling cannot be identified.

For example, in [9], depression diagnosis was investigated using facial dynamic analysis, where extracted features were transformed using sparse coding. Sparse coding aims at reducing the complexity of dynamic features and aims at suppressing the noise in a feature. Sparse coding is a compression method similar to feature transformation, where the original features after the coding cannot be identified. Likewise, in [10], where they used Principal Component Analysis (PCA), a feature transformation method, for dimensionality reduction of acoustical and perceptual speech features to detect depression. In [11], non-linear down-sampling function for dimensionality reduction was applied on speech. Even though [12] employed both transformation method and a voted version of correlation-based as a filter feature selection method, the selected features based on the filter method were not reported.

Sharing a similar goal of model interpretation as this current work, [13] recently developed a method to measure depression severity automatically using face and head modalities. The facial features (shape representation) were isolated from head movement

using barycentric coordinates. They also extract head movement, where they used feature reduction on the two modalities such as PCA and mRMR (minimum Redundancy Maximum Relevance). The face and head movement were presented as a histogram to interpret the results, where the velocity features of facial shape representation show strong discrimination power with respect to depression severity. To further this line of research on interpretability, we analyse depression modelling from features extracted from speech behaviour, speech prosody, eye activity as well as head movement in a multimodal manner.

3 Background on Feature Selection

In general, feature selection methods are categorised to supervised, when the classes' labels are known and used to evaluate the features, and unsupervised, when no labels are available and features are clustered and evaluated for redundancy. This work utilises supervised feature selection methods only, since the focus here is supervised depression detection modelling. Supervised feature selection techniques aim at finding small and representative features that differentiate classes from each other. This is done by removing redundant features and features that do not add value to distinguish the classes.

There are several categorisations for feature selection methods, which are based on their applications, input and output data types (see Table 1). For example, feature selection methods that evaluate all extracted features at once are under flat (static) feature category, while feature selection methods that evaluate each extracted feature as they come available (e.g. online streaming) are under dynamic feature category. Flat feature selection methods are traditionally employed in the literature, where their algorithms assumes that the features are known and extracted in advance. Dynamic feature selection methods are excellent when the total number of features are not known in advance, and they evaluate the new feature to decide (depending on the algorithm) to either include it, replace an old feature with the new one or ignore it. Even though the feature space in our investigation is known in advance, we also chose to apply the dynamic feature selection as an approach to analyse the selected features by these methods. Another family of feature selection methods deals with structural data, where relations between features are learned to construct a structure (e.g. tree), and then the selected features will be based on their location in the structure (e.g. a feature that is located in higher nodes of a tree are more likely to be selected than one in an isolated branch of the tree).

Regardless of the categorisations, every feature selection method uses a different technique/algorithm to evaluate the features. Some feature selection techniques evaluate each feature individually and their contribution in distinguishing the classes without considering the relationship with other features (e.g. t-score). Other techniques evaluate the correlation of features with the class and each other (e.g. Correlation Feature Selection (CFS)). Some techniques evaluate feature groups instead of individual features, where the feature group as a whole is evaluated for selection.

Moreover, feature selection techniques/algorithms differ in the type of input features to be evaluated, where some methods only evaluate discrete input data, while others evaluate both continuous and discrete input data. Based on the output of the method, feature selection methods can be categorised to ranking features, scoring features and selecting feature subset. Ranking methods sort the features based on their importance and value in distinguishing the classes. Scoring methods output a score for each feature representing their importance, such that the higher the feature's score the higher its value in separating the classes. On the other hand, methods that output a feature subset do not give a score or a ranking to individual features, rather they produce a subset of features that performs better in classifying the classes.

For dynamic feature selection, Scalable and Accurate Online Feature Selection (SAOLA) [14] and group-SAOLA [15] were proposed as an online pairwise comparison with a focus on scalability solutions. Relevancy and redundancy of features are analysed using mutual information for the decision of including or excluding them. A continuous comparison between the previously selected features and the newly arrived feature, where the feature with lower relevance with the target class is removed. In the group-SAOLA algorithm, features are divided into groups (e.g. for image analysis, feature groups like SIFT features and colour and shape features [15]), and within each group, several individual features exist. The individual features in a group are analysed first, where redundant and irrelevant features within that group are removed. Then the remaining features from different groups are analysed, where redundant features from other groups are removed. Online Streaming Feature Selection (OSFS) [16] and fast-OSFS [17] also use mutual information in verifying the relevance and redundancy of features. Alpha-investing is a streaming feature selection method that uses statistical criterion to analyse the relevancy of newly arrived features [18]. Alpha-investing dynamically adjusts a threshold on the p-statistic to controls adding a new feature to the model. That is, if the p-value of a new feature is greater than the threshold, then the feature is added to the model. In Alpha-investing, OSFS and fast-OSFS methods, once a feature is selected it will not be removed. Instead, the new arriving feature will be evaluated for inclusion based on its relevance to the class and its redundancy to already selected features. The above dynamic methods use a α-statistical significance level to determine the inclusion of a new feature. Alpha-investing uses dynamically adjusted α, while the others use a predefined α for inclusion, where in this work it is set to 0.1. Adaptive group LASSO (AGLasso) (least absolute shrinkage and selection operator) was proposed to overcome inconsistency in selecting features from LASSO and group LASSO. The LASSO method applies a shrinking (regularisation) process by penalising the coefficients of the variables, where only the variables with non-zero coefficient are selected. LASSO uses the l_1 penalised least squares criterion to evaluate features, which is the sum of the coefficients' absolute values. As a result, many coefficients will be zeroed under LASSO with high values of a threshold (selected in advance). Adaptive group LASSO has flexibility through weighting each coefficient differently to avoid applying the same penalty, which could result in over and insufficiently shrinking regression coefficients.

Some feature selection methods learn a structure from the analysed features to select the best features, including network, tree, and graph structures, as well as rough sets. Unlike most traditional methods that are limited to find interactions between few variables, structure data methods capture high-order interactions between the variables. To learn a network structure from input features, Bayesian Network using Markov Blanket (MB) is widely used. We chose two methods from this group namely the traditional max-min parent children MB (MM-MB) [19] and state-of-the-art statistically equivalent signature (SES-MB) [20].

The main difference between MM-MB and SES-MB is that the latter extends the first by finding multiple subsets of feature that have statistically equivalent performances. Both methods utilise Bayesian networks as graphical models in order to give compact representations of multivariate distributions. The graph composed of nodes that represent variables, and edges that represent relations between the variables, either parent or child. MB is derived based on the parent, children, and any additional parent of children (spouses) of a node. MB of a feature finds redundant features to be eliminated, while MB of a target class comprises a set of selected features. In SES-MB, it is assumed that multiple MBs exists for a target class, where the best representative set is determined for final selection.

Tree structure using random forest variations has been also utilised for feature selection, where five methods were used in this work. A random forest (RF) is used to measure feature importance [21], where the features with the highest importance scores are selected individually. However, this approach does not consider feature redundancy. Regularised RF (RRF) recursively split data, then penalise selecting a new feature that has similar gain (e.g. information gain) to the features used in previous splits [22]. Guided RRF (GRRF) method employs ordinary RF to guide RRF for selecting the features based on feature importance scores [23]. Selection of grouped variables using random forests was proposed (GRF), where permutation-based importance measure is used for each feature group [24]. A most recent approach used Gradient Boosting for feature selection [25], where sequential trees are used for learning and regularised by LASSO.

Graph-based feature selection offers similar benefits of structured data, where the problem feature selection is mapped to an affinity graph (features are the nodes). Infinite Feature Selection (Inf-FS) finds a path in the graph that connects a subset of features to evaluate the importance of each feature while considering all the possible subsets of features [26]. Infinite Latent Feature Selection was proposed as an extension of Inf-FS, where relevancy is modelled as a latent variable [27]. Similarly, Eigenvector Centrality (EigenC) assesses the importance of a feature based on its centrality and the importance of its neighbours [28].

Rough set theory (RST) has been used for feature selection for its ability to deal with incomplete knowledge. In feature selection, RST finds a reduct set of attributes, which is a set of features that have high accuracy in classification. Several algorithms are used to find the reduct. Quick Reduct is a well-known method that uses a greedy search algorithm to select a subset of features using dependency degree as stopping criteria [29]. Dynamically Adjusted Approximate Reducts (DAAR) modifies Quick Reduct method by including an additional stop condition, which is a random permutation [30]. The near-optimal (nearOpt) implements fast heuristic algorithms to obtain one near-optimal attribute reduction instead of finding all reducts [31].

Traditional feature selection assumes a flat (static) and known in advance predictors. Flat feature selection are categorised to filter, embedded and wrappers, where the well-known and widely used methods in the literature are employed in this work. Filter methods use statistical (e.g. t-score [32], Chi square [33], CFS [34]), similarity (e.g. Fisher score [35], ReliefF [36], Spectral Feature Selection (SPEC) [37]) and information theory approaches (e.g. mRMR [38], Joint Mutual Information (JMI) [39], Conditional Mutual Information Maximisation (CMIM) [40], Double Input Symmetrical Relevance (DISR) [41]) to select the best features. While embedded (e.g. LASSO [42], L1-SVM [43], Elastic Nets [44], Ridge [45]) and wrappers (e.g. genetic algorithms (GA) [46], Boruta [47], conditional covariance minimisation (CCM) [48], recursive feature elimination with a linear SVM (SVM-RFE) [49], SVM-Backward feature selection [50]) employs classification algorithms for evaluations.

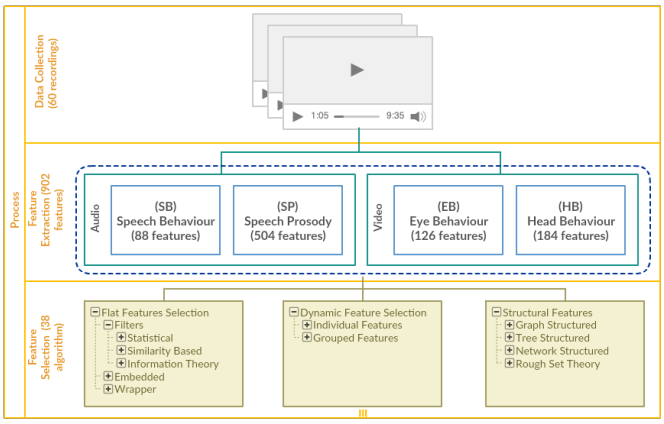


Fig. 1: Preparation Phase of the Feature Selection Framework

4 Selecting Feature Framework

In this section, we introduce our proposed feature selection framework for the interpretation and classification of depression detection model. Our proposed framework for feature selection comprise of three main phases, preparation phase (see Figure 1), feature selection process phase (see Algorithm 1), and aggregation phase (see Figure 3).

4.1 Preparation Phase

The steps in the preparation phase start with collecting video recording of subjects (depressed and control). The raw video recording is analysed to extract several behavioural features from different modalities. A variety of feature selection algorithms from different families are implemented to be applied to the extracted features.

4.1.1 Depression Datasets

The main dataset used in this work is a real-world data collected in an ongoing study at the Black Dog Institute, as detailed in [51]. For the generalisation purposes, we used two other depression datasets, University of Pittsburgh depression dataset (Pitt) [52] and Audio/Visual Emotion Challenge Depression Dataset (AVEC) [53].

BlackDog: The Black Dog Institute is a clinical research facility in Sydney, Australia, offering specialist expertise in depressive disorders. Only subjects who fit the criteria of healthy controls, as well as depressed patients, are included. All depressed subjects met the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders - fourth edition) criteria for either moderate or severe depression. Quick Inventory of Depressive Symptomatology self-report (QIDS-SR) [54] was used to score the severity of depression, (QIDS-SR score of 11-15 points refer to a "Moderate" level, 16-20 points to a "Severe" level, and $\geq$ 21 points to a "Very Severe" level). Control participants were also screened for history of psychiatric and neurological illness. Once subjects were found to meet the inclusion criteria, they are invited to undergo the experimental paradigm.

Participants, both depressed and control, are audio-video recorded in one session only. The audio-video experimental paradigm contains several parts, including answering open-ended questions. The interview is conducted by asking specific open-ended questions, where the subjects are asked to describe events in their life that had aroused significant emotions. This item is designed to elicit spontaneous, self-directed speech and related facial expressions, as well as overall body language.

In this work, a gender-balanced subset of 30 depressed subjects and 30 controls were used for the analysis. For depressed subjects, the level of depression was a selection criterion, with a mean of 19 points (range 14-26 points) of the diagnoses using QIDS-SR scores. Even though the sample size used here is relatively small, this is a common problem in similar clinically validated datasets. Furthermore, from the parts of the experimental paradigm, only the interview section was analysed for feature extraction.

Pitt: The data are from a subset of 57 participants in a clinical trial for treatment of depression conducted at the University of Pittsburgh Medical Center. Treatment consisted of antidepressant medication (i.e., an SSRI) or interpersonal psychotherapy [52]. Both are evidence-based treatments for depression.

All participants met DSM-IV criteria for Major Depressive Disorder (MDD) at start of the study. Audio-video recordings were obtained during depression severity interviews at 7-week intervals over 21 weeks beginning at week one. HRSD (Hamilton Rating Scale for Depression) scores of 15 or higher are generally considered to indicate moderate to severe depression; and scores of 7 or lower to indicate a return to normal [55].

Nineteen participants scored 7 or below at one or more sessions. For inclusion in the study, we randomly sampled one low-depression session from each of these participants. We then randomly sampled a session rated as severe from among an equal number of randomly selected participants that scored in the severe range. Thus, the final sample consisted of 38 participants: 19 with a low-depression session and 19 with a severe-depression session.

AVEC: The Audio/Visual Emotion Challenge (AVEC) is a subset of the German audio-video depressive language corpus (AVDLC). In its 2013/2014 versions, AVEC included a challenge on an automatic estimation of depression level [53]. The database includes 340 video clips of 292 subjects, with only one person per clip, i.e. some subjects feature in more than one clip. The speakers were recorded between one and four times, with a period of two weeks between the measurements. However, in this work we only select one session per subject, where the subjects from the severe and low depression do not cross.

In AVEC dataset, the depression severity is based on the Beck Depression Index (BDI), which is a self-reported 21 multiple choice inventory [56]. The BDI scores range from 0 to 63, where 0-13 indicates minimal depression, 14-19 indicates mild depression, 20-28 indicates moderate depression, and 29-63: indicates severe depression. The average BDI-level in the AVEC dataset was 15 points (standard deviations = 12.3).

The AVEC depression database contains naturalistic video and audio of participants partaking in a human-computer interaction experiment guided by PowerPoint and contains several tasks including telling a story from the subject's own past (i.e. best gift ever and sad event in childhood).

In this paper, a balanced subset of AVEC database is selected based on the BDI score. We categorised the recordings in binary groups indicating severe-depressed where BDI score is more than 29, and minimal-depressed where BDI score is less than 13. Since there were only 16 subjects with a BDI score more than 29, the same number of subjects is selected with ascending low BDI score from 0 to 4. The spontaneous childhood storytelling from the recording tasks is analysed for feature extraction, in order to match the spontaneous interview from BlackDog and Pitt datasets.

As can be seen, the datasets differ in aim, depression assessment/scoring method, and recording procedure. BlackDog aims to compare depressed patients with healthy controls, while both Pitt and AVEC datasets measure and monitor depression severity. Each dataset used different depression assessment tools, which made the depression scores incomparable. In BlackDog and Pitt datasets, all depressed subjects met DSM-IV criteria initially. In BlackDog, depression was assessed using QIDS-SR from both depressed and control subjects. In Pitt, depressed subjects were classified by their subsequent score on a depression severity interview (HRSD, which is the gold standard in clinical trials). In AVEC, all subjects were assessed using a cut-score on the self-report BDI. These methods correspond to the various ways that depression is assessed in research and clinical practice. For consistency across the three datasets, each of the respective scores was converted to its QIDS-SR equivalent using the conversion table from [57]. Each dataset is treated and classified as a binary classification task. That is, with BlackDog dataset the system classifies depressed from control subjects, while with Pitt and AVEC the system classifies severe depression from low depression.

4.1.2 Modalities and Feature Extraction

Behavioural patterns (using statistical measures) of subjects' responses during the interview (interaction) were extracted from different modalities, which are; speech behaviour, speech prosody, eye activity, and head movement. Given the differences not only between subjects, but also between dataset recording environments, normalising each extracted feature was performed within-subject session as described below, and listed in Table 4.

Speech behaviour (SB) (for BlackDog only) Verbal cues and interaction style observed during clinical interviews showed significant differences between depressed and control subjects [58]. Recently, [59] found that speech behaviour features (e.g. speakers' turns, laughter) performed better than other modalities' features (e.g. speech prosody, visual features) in depression severity modelling. The speech interaction pattern during interviews is extracted, following the methodology in [51], (which was done for the BlackDog dataset only due to differences on the other datasets). For the BlackDog dataset, the interviews were manually labelled to separate speakers (i.e., research assistant (RA) and the subject) and to separate several parts of the speech signal for analysis. A total of 88 speech behaviour features are extracted, where these features are grouped listed below. For each feature group, 9 statistical measures are calculated including the average, maximum, minimum, range, variance, standard deviation, total, rate, and frequency of occurrence.

Speech prosody (SP): A recent review on the utilisation of speech prosody features in modelling depression detection and depression severity showed a great impact of these features in the accuracy of the model [3]. Statistical functionals from low-level prosody features were extracted from sounding segments, following the procedure in [60], where the raw features are extracted with frame size set to 25ms at a shift of 10ms and using a Hamming window. The most common features in the depression detection literature from the fields of psychology and affective computing were extracted as listed in the feature groups

below. Following the literature, and to increase the accuracy of depression detection, the first (Δ) and second ($\Delta\Delta$) derivatives of each low-level feature were also extracted [3]. Then a total of 504 statistical functional features are calculated in session-level, where 6 statistical measures are extracted for each feature and its Δs derivatives that include mean, minimum, maximum, range, variance and standard deviation.

Eye activity (EB): Eye movements of depressed patient studies were reviewed in [61], where such features were shown to have statistical discriminating power between patients with depressive disorders from controls. Such features were also used for modelling depression detection, where they show promising results [62]. Following the methodology of eye feature extraction in [63], eye activity (e.g. blinking, iris movement) were extracted using eye detection and tracking model, for each eye in each frame (25 fps). A total of 126 statistical features (average, maximum, minimum, variance, and standard deviation) of the feature and its Δs derivatives were extracted and grouped as listed in Table 4.

Head movement (HB): Studies on depressed patients' behaviour demonstrated a pronounced nonverbal behaviour including head movement that reflects depression persistence [64]. Modelling depression detection using head gesture and movement showed supporting results to other cues [65]. Following the procedure in [63], we extract 3 degrees of freedom head movement behavioural patterns from each frame (25 fps). From these pose features, as well as their velocity and acceleration (Δs derivatives), we extract a total of 184 statistical features as grouped below. The statistical measures include maximum, minimum, range, mean, variance, and standard deviation for the features, its derivatives, and their changes, as well as maximum, minimum, range, average, and rate of head direction duration

4.2 Feature Selection Process Phase

In this work, as described in Section 3, we select a variety of feature selection techniques from each category, which are listed in Table 1. To ensure fair comparison and evaluation of the implemented feature selection techniques, all features with continuous data are converted to discrete data. Moreover, feature selection methods that output a score for each feature could be converted to ranking, and ranking could be used to select a feature subset using a specific threshold. However, the opposite is not always possible. Therefore, in order to aggregate the results of the feature selection methods, feature selection output was used to select a subset of features using a threshold of 10%, as detailed in Section 4.3. We acknowledge that some feature selection methods are sensitive to the training sample and to the order of features in the input data. However, we hypothesise that the features that are selected by these methods will be either strengthened if commonly selected by other feature selection methods, or weakened if not commonly selected by other feature selection methods during the aggregation phase. That is, the feature aggregation phase ensures a valid selection of features to increase the stability, interpretability, and generalisability of the model, and eliminates randomised factors, sensitivity to training sets, and sensitivity to feature order of the feature selection methods. This is done using ensembles of different splits of training data, in multiple rounds of 38 feature selection methods. Moreover, the methods are evaluated through two different stability measures for aggregation, to ensure only features that are robust to these issues will be selected. Similarly for feature selection methods that select one feature from a group of redundant features.

TABLE 1: Properties of Implemented Feature Selection Methods

Feature Selection method				I	O	G	Ref
Dynamic Features			Group SAOLA	D	R	Yes	[15]
			Adaptive Group LASSO	Sp	Sb	Yes	[66]
			Fast OSFS	C	Sb	No	[17]
			OSFS	C	Sb	No	[16]
			SAOLA	D	Sb	No	[14]
			Alpha-investing	D	Sb	No	[18]
Structured Data	Network		MM-MB	C	Sb	No	[19]
			SES-MB	C	Sb	No	[20]
	Tree		RF	D	R	No	[21]
			Gradient Boosting	D	R	No	[25]
			RRF	D	R	No	[22]
			GRRF	D	R	No	[23]
			GRF	D	GR	Yes	[24]
	Graph		Inf-FS	D	R	No	[26]
			Infinite Latent	D	R	No	[27]
			Eigenvector Centrality	D	R	No	[28]
	Rough Set		Quick Reduct	D	Sc	No	[29]
			DAAR	D	Sc	No	[30]
			nearOpt	D	Sc	No	[31]
Flat Features	Filters	Statistics	t-score	C	Sc	No	[32]
			Chi square	C	Sc	No	[33]
			CFS	C	Sc	No	[34]
		Similarity Based	Fisher Score	C	Sc	No	[35]
			ReliefF	C	Sc	No	[36]
			SPEC	C	Sc	No	[37]
		Information Theory	MRMR	D	Sc	No	[38]
			JMI	D	Sc	No	[39]
			CMIM	D	Sc	No	[40]
			DISR	D	Sc	No	[41]
	Embedded		LASSO	Sp	R	No	[42]
			L1-SVM	Sp	R	No	[43]
			Elastic Nets	Sp	R	No	[44]
			Ridge	Sp	R	No	[45]
	Wrappers		GA	C	Sb	No	[46]
			Boruta	C	R	No	[47]
			CCM	C	Sb	No	[48]
			SVM-RFE	C	Sb	No	[49]
			SVM-Backward	C	Sc	No	[50]

I: Input type, **O**: Output type, **G**: Whether the method evaluate features as groups; **D**: Discrete values, **Sp**: Sparse values, **C**: Continuous values, **R**: Rank of features; **Sb**: Subset of features, **GR**: Rank groups of features, **Sc**: Scores of features;

Therefore, ensembles are used. One way of implementing ensembles is by using several feature selection methods on the same data, then aggregate the selected features from these methods. Another way of implementing ensembles is by using the same feature selection method with different parts of the sample data for training (e.g. hold out). In this work, we follow both ensemble methods, in two levels. The first level of ensembles is detailed in Algorithm 1 and described below, while the second level of ensembles and their aggregation is presented in Section 4.3.

For each feature selection method, we follow the same process to reduce variability that might compromise the comparison or the aggregation of the results. Some of the feature selection methods could only apply on discrete data, while others could work on both discrete and continuous data (see Table 1). To equalise the process and the comparisons between the feature selection methods, we discretised the continuous data for all feature selection method. Discretisation is the process of estimating the number of nominal values from continuous sample data, and then categorise each continuous value to its nearest nominal value. Even though advanced methods for discretisation exists, such as k-means clustering, we used Sturge's Rule for its balance between simplicity and avoiding overestimating the number of class intervals (nominal values). This choice is also to avoid overfitting the discretisation process on one dataset since we aim to generalise using different depression datasets. The formula used for discretisation is $K = 1 + \log_2(n)$, where K is the number of class intervals (bins) and n the number of observations in the set. Since we are using three different

Algorithm 1: Process of Feature Selection Methods

Data:
$data$: dataset (samples x extracted features) with labels
M: list of feature selection methods
th: feature cut threshold
Result:
$Sfea$: Array of subset of selected feature from each method
Stb: Array of stability measure for each method

```
for m ∈ M do
    for run ← 1 to Runs do
        for ensemble ← 1 to Ensembles do
            train ← split(data, holdout)   /* random train and
              test subsets split */

            Dtrain = discrete(train.values)          /* convert
              continuous data to discrete */

            fea = m(Dtrain, train.labels, [th])  /* apply method on
              the discretised train subset providing the
              labels for supervised feature selection and
              threshold if required */

            if m is Scoring || m is Ranking
             then  /* method m outputs feature ranking or
              scoring */

                Efea[ensemble] = fea[1...th]    /* select top
                  th features (sorted score/rank) */
            else
                Efea[ensemble] = fea      /* method already
                  returns 1...th feature subset) */

        Rfea[run] = aggr(Efea, th)      /* aggregate selected
          features from all ensembles */

    Stb[m] = stability(Rfea)          /* Calculate stability
      measure between the selected features of all Runs
      */

    Sfea[m] = Intersection(Rfea)      /* Get features that
      intersection between Runs */
return Sfea; Stb;
```

datasets of different sizes, we chose n to be the average sample size between the three datasets. That results in $K = 7$, which was kept constant in all feature selection methods as well as for classification phase for consistency.

Moreover, as discussed earlier, some feature selections rank the features, score them or select a feature set. We cannot convert a selected feature set to scores or ranking, while the opposite is possible. Therefore, all ranking and scores were converted to a feature set by sorting them and then selecting the top feature based on a specified threshold. In this work, the threshold is set to 10% of the total number of the analysed features. We chose 10% to have a small number of features suitable to our datasets size to avoid the curse of dimensionality. This threshold is kept constant in all analysis of the modalities and in the different datasets to assure consistency in the comparison.

Since most feature selection methods are sensitive to the training data, several iterations are performed using random training subsets, then the selected features from each iteration are aggregated. The aggregation function depends on the output of the method. For example, if the method outputs a score, an average score of all iterations is calculated for each feature. Then the average scores are sorted to select the top scored features. In this work, we use 50 iterations (ensembles), where in each ensemble 80% of the data samples are split randomly for feature selection method training. Since the selected features are converted to a subset for all feature selection methods from each ensemble, the

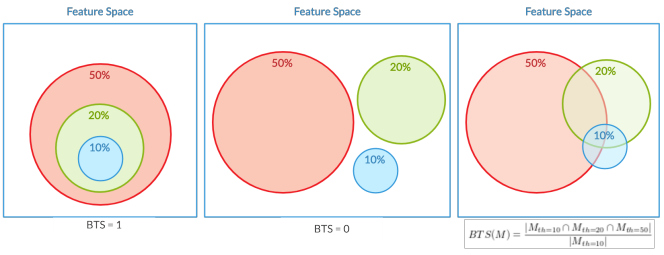


Fig. 2: Between Thresholds Stability Measure (BTS)

aggregation function in this work is the frequency of a feature to be selected by all ensembles. Then the top 10% features with the highest selection frequency are selected. In the case of several features that had the same frequency at the end edge of the top 10%, the selection of the features will be based on the feature ID, where the lower ID is selected first. For example, features 14, 255, 509 all have a frequency of 37, and they are the last to be selected, feature 14 will be in the final selected feature set. This is done instead of a random selection of these features to reduce the variability when aggregating features from all methods.

To calculate the stability of a method, the selected features from separate runs are compared. The aggregated features from the ensembles of each method are run twice to calculate the stability measure between the selected features from the two runs. We use two stability measures for this purpose, the traditional Jaccard similarity index (JI) (also called Tanimoto similarity), which is suitable for comparing sets of single points (in our case the selected features' IDs). Other stability measures exist to compare the scores or ranking of features, which are not used in this work. JI of a method is calculated by the number of overlapping features between runs, over the number of all selected features from both runs, as in the following formula:

$$JI(M_{r1}, M_{r2}) = \frac{|M_{r1} \cap M_{r2}|}{|M_{r1} \cup M_{r2}|}$$

where M_{r1} and M_{r2} are the aggregated selected feature for method M in each of the two runs, respectively. Finally, the aggregated selected features from both runs are intersected to select the strongest features of the method. Therefore, the algorithm concludes by producing the stability measure and the intersected selected features of the runs.

We also propose another stability measure to evaluate the stability of a feature selection method to be used in a combination with Jaccard index. We name the proposed stability measure as Between Thresholds Stability (BTS) (see figure 2). In this stability measure, we consider a method to be fully stable when the top 10% selected features have full overlap with the top 20% selected features as well as in the top 50% selected features. These specific thresholds were selected heuristically, with the rationale of measuring the overlap of top selected features with different levels of flexibility (too lax, too firm, and in between). Similarly, the method is considered not at all stable when there is no overlap between the selected features from different thresholds. Therefore, the stability is measured based on the number of features that overlaps with different increasing thresholds over the total number of features in the smallest threshold, as shown in the following formula:

TABLE 2: Summary of Aggregation Levels

Aggregation	Description
Method-level	Selects the final feature set that is robust to randomness based on stability measures of each feature selection method. Gives an insight into the strongest features that differentiate depressed behaviour in each modality.
Modality-level	Accumulates the strongest features from each modality and the combined modalities using intersect (strict) and union (lenient). Captures both features that interact with other features within individual modality and between different modalities. Mainly meant for modelling depression detection, as well as generalising the selected features on different depression datasets.
Dataset-level	Finds a set of features that could generalise depression modelling regardless of the dataset using relaxed intersection.

$$BTS(M) = \frac{|M_{th=10} \cap M_{th=20} \cap M_{th=50}|}{|M_{th=10}|}$$

where M is the feature selection method, and th is the used threshold. This is performed over the intersected features from the two runs for each threshold level, where each threshold level follows Algorithm 1 process.

Moreover, to evaluate the performance of the feature selected methods they were also compared to a random guess method, where we follow the same process (ensembles and runs) for a random feature selection. Stability measures were calculated in the same way for the random method.

4.3 Aggregation Phase

The process as explained so far, is performed over each feature selection method to select the most promising features in a feature space. This is executed on each modality individually (e.g. SB) and on the combined modalities (All-M). We performed the process over the full feature space from all modalities (All-M) to find the strongest features and their interactions when all modalities are fused. However, this could result in selecting features mostly from the strongest modality without giving a balance to other modalities in selecting their strongest features. This could be avoided by applying the feature selection framework on individual modalities as well. Furthermore, this will be beneficial for generalisation investigation, since not all modalities exist in all depression datasets (i.e. no SB for AVEC dataset procedure).

To select a final feature set for interpretation and modelling, we perform two-stage feature aggregations in each dataset. The first aggregation stage is on the feature selection **Method-level**, while the second stage is on **Modality-level** as described below and summarised in Table 2. For generalisation investigation, **Dataset-level** aggregation is performed to select the common features that are selected from individual datasets.

Method-level Aggregation: The process of methods-level aggregation is illustrated in Figure 3. Basically, the aggregation uses both stability measures, Jaccard similarity index and between thresholds stability, as weights to evaluate each feature selected by each method, then the average weights of each feature are calculated. The average weighted features from each stability measure are then sorted and the top 10% features are selected. Features that have conflicting stability weights are eliminated through an intersection function. The remaining features are the final selected feature for the modality, which is used for the interpretation and modelling of depression recognition.

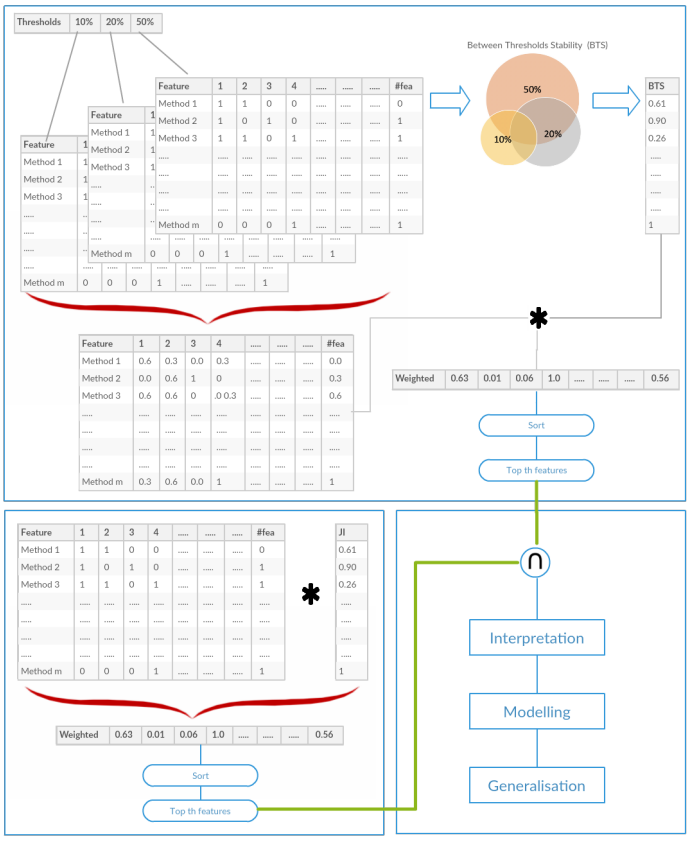


Fig. 3: Method-level Aggregation of Selected Features Using Different Stability Measures for each Modality
(**Top**:The bottom matrix of this top-subfigure is calculated based on the voting of each feature from the top 10%, 20%, and 50% of selected features from each method. Then, the votes of each feature are multiplied by the BST in each method. To aggregate the features from all methods, the average of the weighted votes of each feature is calculated and sorted to select the features with the top 10% highest weights.
Bottom-left: the selected features from top 10% of a method are multiplied by the JI of that method. Then the average weights of each feature are calculated and sorted for the top 10% features to be selected.
Bottom-right: the selected features from both JI and BTS weights are intersected. The final selected features are analysed for interpretation, modelling and generalisation.)

For JI (bottom-left of Figure 3), features that have been selected by a certain method (indicated by 1) are multiplied by the JI of that method. Then the average weights of each feature are calculated and sorted for the top 10% features to be selected.

The process for BTS (top of Figure 3) is slightly different. The feature selection process of algorithm 1 is performed on three thresholds, 10, 20 and 50. BTS is calculated for each method, where the aim is to give higher weight for methods that consistently select the same features when using different thresholds. To aggregate the features from the different thresholds for each method, the frequency of a feature being selected from the three thresholds is calculated. This is to give a higher frequency for features that demonstrate consistently in all thresholds. The frequency of each feature in a method is then multiplied by BTS of that method to produce a weight for each feature in each method. Similar to JI step, the average weight of each feature is used to sort the features to select the strongest 10%.

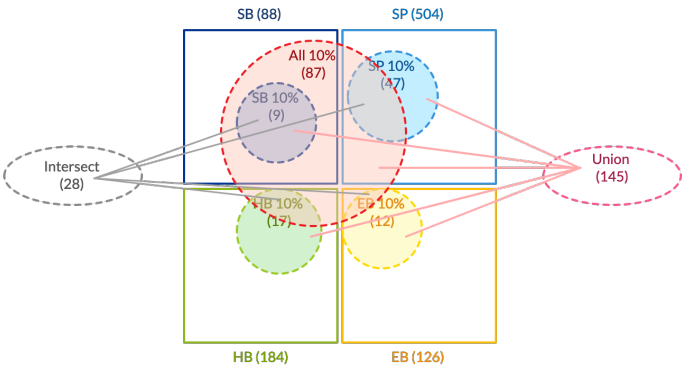


Fig. 4: Modality-level Aggregation of Selected Features using Intersection and Union
(between parentheses is the number of features), **SB**: Speech behaviour, **SP**: Speech Prosody, **HB**: Head behaviour, **EB**: Eye behaviour

As mentioned above, the selected features from both JI and BTS weights are intersected to reduce the features that might have a conflicting weighting from the stability measures. This will result in the final selected features for the modality that have been analysed.

This process ensures that feature selection methods that have high stability measures would get a higher effect in selecting the final feature set. The same applies to feature selection methods that have low stability measures. Methods-level aggregation could also be done by only selecting methods of high stability by disregarding the feature selection method (weight of zero) with stability measures less than a certain level (e.g. stability measures of a method are less than the ones from random selection). We choose not to disregard any feature selection method used in this work, as we believe the method weighting procedure will eliminate the effect of such methods on the overall selected features. This level of aggregation gives an insight into the strongest features that are consistently chosen to differentiate depression in each modality.

Modality-level Aggregation: By method-level aggregation, around 10% of the strongest features from each modality will be systematically selected, which is performed on individual modalities and the combined modalities. The second stage of aggregation is to further refine the feature selection to fuse the strongest features from each modality and the combined modalities for modelling depression detection (see Figure 4). The feature spaces between the individual modalities are totally independent, while each modality shares part of the feature space of the combined modalities. We assume that running the feature selection process on the full feature space, beside the individual modalities, will produce varied features in addition to the ones selected when applied to individual modalities sub-feature spaces. That is, when using full feature space, features that have strong correlation with features from the other modalities will be selected. Moreover, we also expect that there will be bias toward the strongest sub-feature space. Therefore, modality-level aggregation assures that the strongest features are selected from each modality and the combined modalities to capture both features that interact with other features from different modalities, as well as the strongest features from each modality. We perform this aggregation using two methods: intersect, which is strict, where only features from individual modalities that intersect with combined ones are selected, and union, which is lenient, where all features that are selected by individual modalities and the combined ones are selected. This level aggregation is mainly meant for modelling depression detection, as well as generalising the selected features on different depression datasets.

Dataset-level Aggregation: Modality-level aggregation produces feature sets (intersection and union) from individual datasets. To generalise our proposed approach, and find a set of features that have high results for depression modelling regardless of the dataset, we perform dataset-level aggregation. To achieve this, the selected features from individual datasets are aggregated by finding the common features between the datasets. We apply the framework to BlackDog, AVEC, Pitt, and when all datasets are combined. Worth noting that when all datasets are combined, no individual normalisation is performed. That is, the raw values of the features in each dataset are kept as is. Even though the variations between the raw values in each dataset are high given the differences in the recording environment, further normalisation of individual datasets could introduce contamination. Moreover, allowing such variation in raw values would inspect the robustness to recording environment differences of our framework once applied to real-world data. The feature sets from modality-level aggregation for each dataset are intersected. Given the differences between the datasets, we anticipate that strict intersection (agreement of all datasets) will not produce enough features for a generalised modelling of depression. Therefore, we introduce a relaxed intersection, to find the finest feature set that has the ability to generalise to different datasets.

A relaxed intersection is a method used in a variety of applications, such as constraints satisfaction problems such as localisation and control of a robot. To avoid an empty intersection of constraints, relaxed intersection allows relaxation of few constraints. The q-relaxation, denoted as $\cap^{\{q\}} X_i$, of the sets $X_1, ..., X_m$, is the set of all x which belong to all X_i's, except q at most [67]. In q-relaxed intersection, q is set by the application to adjust the level of relaxation. In other words, if q is set to 0, then the result is the strict intersection, where the results are only the items that exist in all sets. While if q is set to the number of sets (m), the results are the union of all items in the set. In this work, we explore intersecting the selected features from different datasets and their effect on depression modelling using different q values for the q-relaxed intersection.

4.4 Modelling Depression Detection

For modelling depression detection, machine learning techniques, such as neural networks (NN), are used. In this work, we use a support vector machine (SVM) to model depression in a binary classification problem (i.e. severe depressed vs. low-/non-depressed). Unlike NN and its deep learning variations that requires huge sample size, SVM is a discriminative method that learns boundaries between classes even when using small datasets, while providing good generalisation properties. SVM has only a few parameters to be tuned (cost and gamma), besides the kernel function, which balances between simplicity and accuracy. These characteristics of SVM makes it suitable for our modelling, especially considering that the modelling is meant to further evaluation of the selected features.

We use supervised subject-independent scenario for all classification tasks. The SVM kernel selected in this work is the radial basis function (RBF), and optimisation for the cost and gamma parameters is performed by a wide range grid search. The range is set to -80 to 80 for cost and 80 to -80 for gamma, with wide

to narrow search steps (40, 20, 10, 5, 2.5, and 0.5). For some sets of features, SVM was not able to find the optimal hyperplane, which might be because the range of these features and their associated labels are too small to be separated. We fixed this for these sets by further adjusting the parameters search by increasing the grid range and adding a fine 0.2 step (these instances are marked with †). Leave-one-subject-out (LOSO) cross-validation without any overlap between training and testing data is used to mitigate for the relatively small number of observations in the datasets. The performance of the modelling is measured in terms of average weighted (balanced) accuracy. For all modelling tasks, features were normalised by discretisation (converting continues data to discrete values). This is performed to be equivalent to the discretisation step in the feature selection process and to minimise the differences between the features in the three used datasets used for generalisation.

For comparisons, the depression recognition is modelled using full feature space for combined modalities and for sub-feature spaces for individual modalities, the final selected features from combined modalities and individual ones, and from random feature selection method. These comparisons will highlight the effect of the selected features in modelling depression recognition.

4.5 Generalisation:

To assess the validity of our proposed feature selection framework, and the robustness to randomness of the selected features in detecting depression, several generalisation approaches are performed. As an initial investigation, we generalise the results of the main depression dataset (BlackDog) on the other two datasets by evaluating the final selected features from the BlackDog dataset, by modelling depression severity detection using Pitt and AVEC depression datasets.

Then we generalise the framework and the selected feature modelling through:

- Applying our proposed feature selection framework on each dataset individually and as combined. This is to investigate the features that are commonly selected by all datasets.
- Modelling depression detection using the selected features from each dataset and the combined datasets on each other. This is to evaluate the effectiveness of the selected features in modelling depression from unseen datasets.
- Performing dataset feature aggregation (Dataset-level) to explore the ability to generalise the final selected features to the datasets. This is done through relaxation of intersection.

5 Applying the Framework on BlackDog Dataset

5.1 Feature Selection Methods Results

Due to the sensitivity of feature selection methods to training sample, ensembles are used to increase the methods' stability, where stability measures evaluate the sensitivity of a method. In this work, we used Jaccard index as a stability measure, since the output of all methods is in the form of a subset of features. We also proposed a new stability measure, which evaluates the selected features from different thresholds, we name it between thresholds stability. Feature selection methods were applied on individual modalities (i.e. SB, SP, EB, HB) and their fusion.

TABLE 3: Stability and Contribution Results of Feature Selection Methods of 10% Threshold

Method	Stability	JI Avg.	JI Std.	BTS Avg.	BTS Std.	Contribution Avg.	Contribution Std.
Dynamic	Group SAOLA	0.61	0.29	0.89	0.15	0.13	0.10
	AGLasso	**0.94**	0.06	**1.00**	0.00	**0.25**	0.21
	Fast OSFS	0.08	0.04	0.28	0.23	0.01	0.02
	OSFS	0.10	0.06	0.39	0.28	0.01	0.01
	SAOLA	0.61	0.31	0.79	0.17	0.09	0.09
	Alpha-investing	0.07	0.03	0.22	0.19	0.04	0.05
Network Structure	MM-MB	**0.64**	0.12	**0.95**	0.05	**0.65**	0.10
	SES-MB	0.61	0.15	0.94	0.06	0.57	0.16
Tree Structure	RF	0.56	0.30	0.94	0.05	0.54	0.18
	RRF	0.37	0.15	0.85	0.10	0.47	0.22
	GRRF	0.59	0.14	**0.97**	0.05	**0.56**	0.10
	GRF	0.53	0.45	0.80	0.45	0.22	0.13
	Gradient Boosting	**0.69**	0.11	**0.97**	0.06	0.47	0.18
Graph Structure	Inf-FS	**0.88**	0.07	**0.97**	0.04	**0.53**	0.05
	Infinite Latent	0.73	0.09	0.93	0.05	0.20	0.13
	EigenC	0.05	0.04	0.06	0.08	0.00	0.01
Rough Set Theory	Quick Reduct	0.46	0.14	0.91	0.09	0.42	0.15
	DAAR	**0.59**	0.21	**0.98**	0.02	**0.50**	0.20
	nearOpt	**0.59**	0.15	0.92	0.06	0.38	0.19
Filters	t-score	**0.93**	0.09	**1.00**	0.00	0.40	0.35
	Chi square	0.83	0.07	**1.00**	0.00	**0.51**	0.14
	CFS	0.74	0.25	0.90	0.14	0.35	0.24
	Fisher Score	0.91	0.06	**1.00**	0.00	0.49	0.25
	ReliefF	0.84	0.10	**1.00**	0.00	0.40	0.19
	SPEC	0.34	0.07	0.74	0.36	0.03	0.04
	MRMR	0.62	0.23	0.86	0.13	0.23	0.22
	JMI	0.77	0.03	**1.00**	0.00	0.40	0.27
	CMIM	0.63	0.18	0.99	0.02	0.33	0.21
	DISR	0.85	0.10	**1.00**	0.00	0.45	0.25
Embedded	LASSO	0.84	0.15	0.97	0.04	0.55	0.13
	L1-SVM	**1.00**	0.00	**1.00**	0.00	0.44	0.18
	Elastic Nets	0.92	0.09	0.97	0.03	**0.58**	0.13
	Ridge	0.78	0.09	0.99	0.02	0.36	0.10
Wrappers	GA	0.45	0.21	0.54	0.25	0.16	0.11
	Boruta	0.79	0.09	1.00	0.00	**0.63**	0.08
	CCM	0.80	0.08	1.00	0.00	0.46	0.11
	SVM-RFE	0.76	0.11	0.99	0.03	0.44	0.09
	SVM-Backward	**1.00**	0.00	**1.00**	0.00	0.06	0.07
Random		0.03	0.04	0.10	0.14	-	-

Results are in term average (avg.) and standard deviation (std.) of individual modalities and their fusion.
***** results in **Bold** are the maximum results in each feature selection group
JI: Jaccard Index, BTS: Between Thresholds Stability, Contribution: is the percentage of features that are included in the final feature set

The aim is to find the strongest features from each modality and to find the features that have a strong interaction between the modalities, respectively. We evaluated the stability of each of the features selection methods in fused and individual modalities. The average (Avg.) and standard deviation (Std.) of JI and BTS stability measures are presented in Table 3.

Moreover, to evaluate the contribution of each feature selection method to the final selected features (listed in Table 4), we count the intersecting features between the ones selected by that method and the final selected features. That is, the overlap between the top 10% features that are selected by each method, and the final aggregated features from all methods are calculated in terms of percentage. The higher the overlap of the features from a single method with the final feature set, the higher the contribution of that method to the overall aggregation. The aim is to evaluate each method contribution against its stability measures. Moreover, the contribution is calculated for fused and individual modalities (see Fig. 4), where the average and standard deviation are shown in Table 3. Comparing the average of JI and BTS of feature selection methods with a random selection method shows variation in the methods' stability. A large standard deviation indicated a high fluctuation of stability between the modalities, where it indicates the sensitivity of the feature selection method to the features of

the analysed modality.

Several feature selection methods performed relatively close to the random method, such as Alpha-investing, Eigenvector Centrality, OSFS, and fast-OSFS. Even though Alpha-investing, OSFS, and fast-OSFS had low stability, the other dynamic feature selection methods (e.g. SAOLA) have high stability performance. That might be due to their approach to removing already added features if a stronger feature is found. Because Alpha-investing, OSFS, and fast-OSFS methods are strict in removing an already selected feature, their stability is affected and their sensitivity to the training sample was pronounced. This indicates that dynamic feature selection methods, in general, are suitable for our dataset and that these specific methods might need tuning for our feature space (e.g. parameterisation). Similarly for the graph-based feature selection methods, where Eigenvector Centrality had low stability compared with the other two methods in the same family.

Illustrated in **Bold**, Table 3 shows the highest stability method in each group. For dynamic features, adaptive group lasso shows high average stability as well as a low fluctuation between modalities compared to other methods in the same group. For the other method groups, most methods performed similarly to the other methods in the same group (except for RRF, SPEC, and GA).

The highest standard deviation of stability was obtained from GRF, which indicates different stability results from each modality. Further investigation showed that its stability from HB modality was equivalent to random level. Since GRF evaluates feature groups, this could indicate that grouping HB features need further tuning. On the other hand, SAOLA, group SAOLA, and CFS had a high standard deviation of stability, where the lowest stability was obtained from using the fusion of all modalities. That might indicate that these methods are sensitive to the size of feature space and their increased feature interaction. Finally, RF high standard deviation of stability is sourced from SP and HB modalities. The common characteristic between these two modalities is that some features are missing a lot of information from most observations (i.e. only 1-2 observations had values for some of the features). Since all methods are exposed to the same features, we believe that this observation highlights the sensitivity of the method to sparse features. This indicates that our proposed method is robust to such sensitivity issues, and is able to eliminate the features that could be selected of methods that are sensitive to the sparsity of the features, the order of the features, training sample, etc.

On the other hand, methods that had the highest stability are L1-SVM and SVM-backward. Closely investigating the selected features from these two methods, we realised that the features being selected are the first 10% of features and the last 10% of features, respectively. This might indicate the inability of the SVM to find any separating hyperplane in these cases, or unsuitability for our discretised features to these methods. SVM based methods could find difficulties in finding optimal hyperplane either because of an inappropriate type of kernel function or because the feature ranges associated with their labels are too small to be separated. These issues can be fixed by changing the kernel function and/or adjusting the parameters to adapt the margin according to the features. In this note, L1-SVM uses a linear regression SVR with l_1 as a penalty, and the SVM-backward method uses a radial basis function (RBF), both with default parameters, which might have affected their performance. On the other hand, SVM-RFE uses a linear kernel SVM, has lower stability compared with SVM-based methods, but produces a variant selected features, which might indicate the suitability of a linear kernel SVM to our features compared to SVR and RBF kernels. Even though parameter optimisation is recommended for SVM based methods, we kept the default parameters for fair comparisons with the other feature selection methods in this work. We also selected several SVM based methods that have different kernels and penalty approaches to have a variety of feature selection methods.

It is expected that a strong feature selection method (one with high stability) would have a higher contribution in the final selected features compared to weak methods. For example, SVM-backward has high stability measures and yet had a very low contribution. Given that the inaccurate results of the SVM-backward, the high stability is mistaken, our framework proved to be robust to false high stability. As can be seen from Table 3, methods that have a high contribution to the final selected features are the ones that have high stability measures and low fluctuations of different modalities.

In summary and to highlight the strongest feature selection methods in our dataset, adaptive group lasso followed by t-score had the highest stability measures, while MM-MB and Boruta had the highest contribution to the final selection. Moreover, most groups of feature selection method were suitable for our dataset, where a few other methods could benefit from parameter optimisation to increase their stability.

5.2 Interpretation of Selected Features

The proposed framework in this work aims at aggregating the selected features from feature selection methods to increase the robustness to randomness of the final selected features. This will not only help in increasing the modelling accuracy, but also will provide an interpretation of the model and its ability to generalise to other depression datasets. The aggregation phase is performed in two levels, method-level, and modalities-level.

The method-level is applied to the selected features from each method for individual and fused modalities. The aim for applying on individual modalities is to find the strongest features from each modality, while the aim for applying on the fused modalities is to find the strongest features with a focus on the interaction between features from different modalities. On the other hand, the modalities-level aggregation is combining the final selected features from each modality and the fused one. This is performed as an intersection (strict), to narrow the features to the distinct ones, and a union (lenient) to capture all distinct features from individual and fused modalities and their interactions. The idea behind this level of aggregation is to find a set of features to be used for modelling depression and test its ability to generalise in different (unseen) depression datasets. The final selected features of both aggregation methods are shown in Table 4.

Applying the proposed framework on fused modalities with a 10% threshold resulted in 87 features. As can be seen in Table 4, most of these features (64 features) are from speech behaviour modality, followed by speech prosody. Since speech behaviour features are expensive to extract (given the annotation effort), it is worth investing in automatic approaches to extract such features. Moreover, this also emphasises the importance of speech interaction in depression detection and their inclusion in any dataset collection procedure. Within the SB features, sounding and silence segments were the weakest features, where only the total duration and number of segments were selected to be informative. This might be due to their redundancy with other features that

TABLE 4: Final Selected Features in each Feature Group for each Modality after Aggregation Phase in BlackDog Dataset

Aggregation			Method-level					Modality-level	
Modality		#	All-M	SB	SP	EB	HB	∩	∪
Feature Groups			(902)	(88)	(504)	(126)	(184)		
SB	First Response	9	9	4				4	9
	Total Response	9	9						9
	Overlap Speech	9	9	3				3	9
	Overlap Laugh	9	9	2				2	9
	RA interaction	9	9						9
	Subject Speech	9	8						8
	Subject Laugh	9	6						6
	Sounding	11	2						2
	Silence	11	2						2
	Speech Rate	3	1						1
SP	F0	18			13				13
	HNR	18			4				4
	Voice Probability	18			3				3
	Jitter	18							
	RMS energy	18							
	Voice Quality	18							
	Log Energy	18			2				2
	Shimmer	18			2				2
	Intensity	18							
	Loudness	18							
	Formants (1-6)	108	11		14			9	16
	MFCC	216	3		10			3	10
EB	Left-Right mov.	30	1			3		1	3
	Up-down mov.	30							
	Openness rate	30							
	Blinking	12	1			3		1	3
	Looking Duration	24				6			6
HB	Yaw mov.	60	4				7	3	8
	Roll mov.	60					4		4
	Pitch mov.	60	3				6	2	7
	Mov. counts	4							
Total Selected features			87	9	48	12	17	28	145

∩: Intersection of features, ∪:Union of features

already extract such features (e.g. subject speech). For SP feature group, formants features (specifically the second derivative $\Delta\Delta$) were the strongest features aside from SB features.

Even though the selected features from fused modalities showed that speech behaviour is the strongest modality, it also shows that feature selection methods are biased towards it. In order to reduce this bias, we applied our feature selection framework to individual modalities to identify their strongest features independently. Having only a 10% threshold, the selected features from individual modalities are listed in Table 4. SB final selected features concentrated on first response and overlap features. Prolonged response to the questions is found in depressed patients, which might be an indication of psychomotor retardation, while less overlap speech is found in depressed subjects as an indication of decreased interaction [58]. In line with the literature [68], SP features showed that fundamental frequency (pitch), formants and MFCC feature groups to be the most informative features for depressed patient speech, which indicates a monotone and psychomotor retardation that lead to a tightening of the vocal tract. HNR features indicating breathiness, voice probability indicating monotone voice, speech energy, and shimmer indicating sounding irregularity were also from the fine selected features. However, voice quality which is another sign of monotonic speech, and jitter which is another measure of sounding irregularity, were not selected, as widely reported in the literature. This might be an indicator of redundancy and a stronger predictor from similar features, such as voice probability and shimmer, respectively.

Even though EB feature group had only two features selected from the fused modalities, given the 10% threshold, a few more features were selected from EB as an individual modality. The two strongest features were the average left-right iris movement and the average blinking duration. These two features are found to be higher in depressed subjects [62]. Duration of looking direction was also found be a strong predictor for depression, where the duration of looking down and left are the main features, which for the depressed participant are signs of avoiding eye contact with the interviewer [62].

HB features have a stronger influence than EB features when selecting features from all fused modalities. Seven features were selected in the fused modalities, where five of which are commonly selected by analysing the HB individual modality. For yaw head movement (nodding no), average duration of looking left as well as features from the speed of movement are the strongest features. For pitch head movement (nodding yes), features from the speed of movement are the strongest features. Similar to yaw and pitch head movement, head roll (tilting) features from the speed of movement are the strongest features. Similar to left eye gazing, head looking left might be an indication of eye contact avoidance, while slow head movements are an indication of psychomotor retardation, which are characteristic behaviour of depressed subjects [65].

Results of modality-level feature aggregation are presented in the last two columns in Table 4. The intersection method aims to find the distinct features commonly selected from individual and fused modalities, which resulted in only 28 features from the original 902. However, this method ignores the interaction between features within the same modality and between modalities. Therefore, a union approach overcomes this issue by capturing such interactions, even though it selects more features than the specified 10% threshold. The aim of this method is to refine the selected features for depression modelling and generalisation as shown in the next sections.

5.3 Effect of Selected Features on Classification

To inspect the effect of the final selected features after the aggregation phase, we model depression detection using (1) all features without selection (All-F), (2) using randomly selected features (following the same process of Algorithm 1) and (3) using the final selected features. The accuracy results are presented in Table 5. Overall, our framework of feature selection showed significant gain in term of classification performance in both method-level and modality-level aggregations compared to random selection and without selection of features.

Even though the paper's aim is to find robust, and discriminative features for depression that can be interpreted and generalised to different datasets, we analyse the classification results and the number of features selected by each feature selection method[1]. This analysis showed a variation of the number of selected features from one method to another, and the classification results. Some single methods outperformed the classification results from the framework selected features. However, their performance was not consistent between modalities. Nonetheless, this analysis could be helpful if the goal is to increase the accuracy of the model.

Using **method-level in the aggregation** phase shows significant improvement in classification results in most modalities compared to using all features (All-F) and randomly selected features, with two exceptions. For EB modality, similar classification results were obtained from both all features (All-F) and the selected features. The second exception is that SP classification results from selected features had a slight improvement compared

1. https://sites.google.com/view/featureselectionframework

TABLE 5: Classification Results using Different Aggregations of Feature Selection on the BlackDog Dataset

Modality/ Features	All-M	SB	SP	EB	HB	⋆∪	∩	∪
All-F	75.00	76.67	60.00	65.00	61.67†	-	-	-
Random	75.00	66.67	76.67	38.33	48.33	-	-	-
Selected	86.67	81.67	78.33	65.00	56.67	81.67	**85.00**	78.33

⋆∪: Features Union from individual modalities (excluding combined modality),
∩: Features Intersection from all modalities, ∪: Features Union from all modalities
†: A fine (0.2) steps grid search was used to find the best parameters

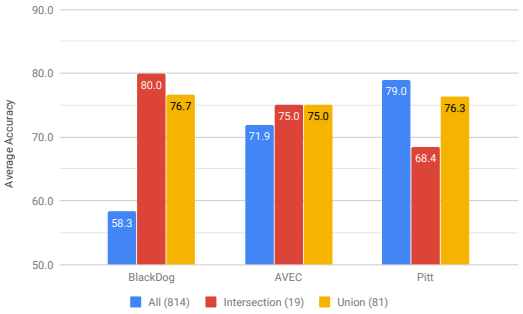


Fig. 5: Generalisation Results using Final Selected Features from BlackDog Dataset (without SB) to other Depression Datasets (the number of features used for modelling are between parentheses))

to randomly selected features. Surprisingly, a significant improvement in classification results was obtained from randomly selected features of SP compared to when using all features (All-F). For HB, the optimal hyperplane was not reached from the original grid search, hence the need to further refine the SVM parameters for HB features. This might be because of the sparse nature of some of the features in SP and HB (i.e. only 1-2 observations had values for some of the features. In the case of SP, randomly selected features might have the effect of removing some of these features that caused the drop in recognition rate. Nonetheless, the constant improvement from the selected features indicates the success of our framework to select the strongest predictors of depression even with the small number of selected features (10% of total features).

Modality-level in the aggregation phase also shows improvement compared to when using all features (All-F). A classification of the union of features selected from the individual modalities (⋆∪) was also explored to be compared with the classification results from selected features from fused modalities (All-M). The union of individual modalities features (⋆∪) performed as high as the highest modality (81.67%), which is the SB modality in terms of the classification results. This is also significantly higher than using all (All-F) and random features from fused (All-M) and individual modalities. This is another indication of the importance of capturing the speech interaction in depressed patients. Nonetheless, as indicated, a slight reduction in the classification results in the feature union from the individual modalities (⋆∪) compared with the case of using selected features from fused modalities (Selected All-M). This might indicate a loss in some features that have a strong interaction with other features in other modalities.

Selecting only features that are commonly selected in individual modalities and the fused case (∩), results in significant improvement in classification results compared to individual modalities. However, this also results in a slight reduction from using selected features from fused modalities (Selected All-M). Given that the final selected features from the intersection method (∩) are only 32% (28 of 87) of the selected features from the fused modalities (Selected All-M), the reduction in classification results is acceptable.

By performing a union function on features selected from the fused modalities and the individual ones (∪), we aimed at capturing both the interaction between features from different modalities with each other, and the strongest features in each modality. Even though the classification results using the union of selected features (∪) performed better than all ((All-F All-M)) and randomly selected features (Random All-M), it had a lower performance than selected features from fused modalities (Selected All-M) and the intersection of selected features (∩), which was unanticipated. The number of features selected in the union features (∩) is 145 features, while it is 87 for the (selected All-M). Inspecting the features that have been included in the (∩) set showed that 36, 10 and 12 features are from SP, EB, and HB modalities respectively. Given that the majority of the added features are from SP, the classification results of the (∩) set had the same accuracy. This could indicate that the model was skewed to the dominated features from the SP, compared to the somewhat balanced modalities from the other feature sets.

Finding the best level of feature intersection (relaxation of intersection) could help balance this issue to improve the classification results while selecting the strongest features and their interaction. However, since the features from individual modalities are from entirely separate feature spaces, a relaxation of intersection cannot be performed in this case. Adjusting the feature selection framework to include feature ranking even for methods that produce feature set, could help further evaluation of each feature to refine the selection.

5.4 Generalising BlackDog Selected Features

It could be argued that the selected features from BlackDog are trained on the same dataset (seen data), which could introduce contamination of the classification results. We acknowledge this risk, however, the sample size in BlackDog would not allow for a validation set. To test the generalisability of the final selected features from the BlackDog dataset, we model depression severity detection (high/low depression) on AVEC and Pitt datasets. Since AVEC dataset collection procedure was based on computer-human interaction, there are no speech behaviour interaction features to be extracted. Therefore, for both AVEC and Pitt only the selected features from SP, HB, and EB are used for the modelling and compared when using all features. For BlackDog dataset, re-modelling was performed excluding SB modality for comparison. The generalisation results are illustrated in Figure 5.

In the BlackDog dataset, the results show a reduction in classification results compared to when including SB (Table 5), which emphasises the importance of SB features in modelling depression. The low classification results from the BlackDog dataset when SB features are removed might be due to the observed sparse features in SP and HB, as mentioned earlier, which hindered the SVM algorithm from finding the optimal hyperplane. Nonetheless, a similar pattern of classification results was observed when using the intersection and union of selected features, where intersection performed slightly better than the union.

Unlike the BlackDog dataset, using all features for modelling depression detection in AVEC and Pitt datasets achieve high classification results. The reason behind this might be that AVEC

and Pitt datasets did not have the same sparsity observation for the same features as in the BlackDog dataset, which could be due to the differences in the datasets signals, protocols, devices, etc. For AVEC, both the intersection and union of selected features performed similarly, which was slightly higher than when using all features. This indicates that the selected features from our proposed framework generalised to AVEC without losing discriminative depression features. On the other hand, for the Pitt dataset, the intersection and union selected features could not out-perform the classification results of using all features. One explanation might be the differences between the datasets' environment and procedure, even though we have taken all possible measures to normalise for these differences. In particular, the subjects with low depression scores in the Pitt dataset were professionally diagnosed with scores ranging between 1-5 (average=2), while the self-rated AVEC low depression score ranges between 0-2 (average=1), and it is zero for the BlackDog dataset (healthy control). The differences between these diagnoses could be another reason that the selected features did not have a similar generalisation power in Pitt compared to AVEC. Nonetheless, the difference in classification results between using the full feature space (814 features) and the union features (81 features) in the Pitt dataset is not significant (2.7% absolute).

On a positive note, the results from AVEC and Pitt in regard to selected union features from different modalities indicate that the features selected from the BlackDog dataset were able to generalise. Moreover, comparing the intersection and union results from the three datasets, the union selected features had a higher ability to generalise with satisfying performance on unseen datasets, given the reduction in the number of features. This confirms that the selected features have a distinguishing strength to detect depression.

TABLE 6: Selected Features using our Framework on Different Depression Datasets

Feature Group			Modality-Level Aggregation							
			Intersect Methods				Union Methods			
	Dataset	#	BD	AV	Pitt	AD	BD	AV	Pitt	AD
SP	F0	18	11	16	2	5	16	18	10	15
	HNR	18	3	1	4	4	6	10	14	6
	Voice Probability	18		2		3	3	9	5	5
	Jitter	18			1				1	
	RMS energy	18								
	Voice Quality	18		6	1	1		6	1	3
	Log Energy	18	1				2			
	Shimmer	18	2		3	2	3	2	7	2
	Intensity	18							1	
	Loudness	18								
	Formants (1-6)	108	14	15		8	23	22	5	10
	MFCC	216	9	5	9	15	10	9	23	20
EB	Left-Right mov.	30	2	1	3	3	3	1	5	3
	Up-down mov.	30			1		1	2	2	2
	Openness rate	30		5	2	1	2	6	2	1
	Blinking	12	3		1	1	5		1	2
	Looking Durations	24	5	1	5	7	6	3	6	9
HB	Yaw mov.	60	3	1	4	3	7	5	8	5
	Roll mov.	60			4	2	4	7	12	9
	Pitch mov.	60	2		2	3	6	5	7	5
	Mov. count totals	4			1				3	
	Total #features	814	55	53	43	58	97	105	113	97

*BD: BlackDog dataset, AV: AVEC dataset, AD: All Datasets combined

6 Generalising the Framework

Following the success of generalising the results of our feature selection framework from BlackDog dataset to other depression datasets, we chose to investigate the generalisation ability from other datasets to each other. We perform this through running the same procedure on individual datasets and when combined to select the strongest features. The selected features from each dataset are then used for modelling depression in each dataset individually. Finally, we perform a dataset-level aggregation using relaxation of intersection to explore the level of feature intersection that would produce the good classification results on all the three datasets and when combined.

6.1 Datasets Selected Features

Applying our proposed feature selection framework on different datasets results in selecting features from each dataset (see Table 6). For BlackDog dataset, we have re-run the framework excluding SB modality to have fair comparisons with the other datasets. As with BlackDog dataset, two approaches of modality-level aggregations are performed and compared.

Intersect approach is a strict method to aggregate the features selected from different modalities. However, this strictness shows the distinct features. Applying this method in different datasets, show that F0, HNR, MFCC, left-right eye movement, eye gaze direction, and yaw head movement feature groups are the most commonly selected groups. In this method, up-down eye movement, head movement frequency, intensity, loudness, jitter, and energy feature groups were not selected in the majority of the datasets, which indicates either a weak contribution in depression or redundancy with other already selected features. Differences between the datasets exist in selected feature groups in the intersection method. The feature groups that were not selected for BlackDog dataset, voice quality and eye openness rate, while for AVEC they are shimmer, eye blinking, and pitch head movement, and for Pitt it is formants.

Union approach of the modality-level aggregation is more lenient than the intersection approach, where all features that have been selected from individual and fused modalities are selected. Commonly non-selected feature groups between the datasets are the same as in the intersection approach. However, the commonly selected feature groups between the datasets included extra groups. Specifically, formants features from Pitt dataset were included in this method. It was unexpected that none of the formants features were selected in Pitt in the intersection method since formants have been widely reported in the literature for their importance. Therefore, their inclusion in the union method is satisfactory, even though only a few features were included.

As can be seen from the pattern of selected features from the different datasets in Table 6, our framework demonstrates a generalisation potential. When applying the framework to the combined datasets, most feature groups were selected, except for the ones that were commonly not selected by individual datasets. This shows that the proposed framework is robust to randomness and able to capture the distinguishing features even with the variations in the datasets.

6.2 Modelling and Generalising Selected Features

The selected features from one dataset were used to model depression severity detection on other datasets, to examine the generalisability of the selected features and the framework. The classification results using features selected from each dataset on other datasets are illustrated in Table 7.

Since the framework is re-applied on BlackDog dataset without SB modality, reduction in classification results is observed,

TABLE 7: Classification Results for each Datasets based on Features Selected from Different Datasets

Classification On		Features Selected From							
		BlackDog		AVEC		Pitt		All Datasets	
Dataset	All-F (814)	∩ (55)	∪ (97)	∩ (53)	∪ (105)	∩ (43)	∪ (113)	∩ (58)	∪ (97)
BlackDog	58.3	81.7	76.7	71.7	61.7	56.7	65.0	**76.7**	75.0
AVEC	71.9	71.9	75.0	87.5	84.4	68.8	65.6	71.9	**78.1**
Pitt	78.9	73.7	76.3	65.8	71.1	84.2	86.8	76.3	**84.2**
All Datasets	**74.6**	71.5	73.8	65.4	73.1	69.2†	70.8†	76.2	75.4†
Average	70.9	74.7	**75.5**	72.6	72.5	69.7	72.0	75.3	78.2

∩: Features Intersection, ∪: Features Union, All-F: All extracted features
(the number of selected features is between parenthesis)

emphasising the power SB features have in detecting depression. Using the features selected from BlackDog to model depression detection on the unseen datasets, a satisfactory performance was seen compared with using all features given the huge reduction in feature space. Modelling using AVEC selected features outperformed modelling using all features in BlackDog and AVEC, but not in Pitt and combined datasets. At the same time, modelling using Pitt selected features could not generalise to other datasets. This might be due to the difference in Pitt procedure, where the subjects are in a treatment session with their therapist, while in AVEC and BlackDog they interact with a computer and a person they have just met, respectively. Therefore, the behaviour and interaction of the subjects in Pitt might significantly differ from that in BlackDog and AVEC. Moreover, the high and low depression range in Pitt subjects is higher than the ones in BlackDog and AVEC, which might influence the intensity of subjects' behaviour. The selected features from the combined datasets were able to generalise on the other datasets and outperform using all features. As expected, features selected from one dataset are able to outperform using all features in the dataset itself. When using the combined datasets, the selected features were able to generalise to all individual datasets, including the combined one. It might be argued that this is caused by overfitting the selected features of a dataset to model depression on the same dataset. However, we believe that the use of random split for training and the use of ensembles have reduced the effect of overfitting, especially considering that some selected features applied on unseen datasets were able to generalise adequately (e.g. BlackDog features on modelling AVEC and vice versa). Further investigation of modelling selected features from two datasets on the third could verify this aspect, which will be done in future work. Comparing the generalisability of the intersection and union features, on average union features has slightly higher classification results than the intersection features. This could indicate that the flexibility of the union approach on including relevant features increases their ability to generalise to other unseen datasets even with a huge reduction in the feature space. This leads to the next section, where we investigate the effect of the level of flexibility on datasets generalisability.

6.3 Dataset-Level Feature Aggregation

After applying the feature selection framework to individual and combined datasets, different features were selected from each feature group. In dataset-level feature aggregation, we aim to further refine the selected features and confirm their ability to generalise to different datasets regardless of their variation. For this level of aggregation, we employ q-relaxed intersection, with different values for q to explore the finest feature set that has

TABLE 8: Selected Features from Different Depression Datasets using Dataset-Level Aggregation

Feature Group			Dataset-Level Aggregation							
			Intersect Features				Union Features			
Relaxation of Intersection		#	$q3$	$q2$	$q1$	$q0$	$q3$	$q2$	$q1$	$q0$
SP	F0	18	17	14	3		18	17	15	9
	HNR	18	6	4	2		17	13	5	1
	Voice Probability	18	4	1			12	6	3	1
	Jitter	18	1				1			
	RMS energy	18								
	Voice Quality	18	6	2			6	4		
	Log Energy	18	1				2			
	Shimmer	18	5	2	1		9	3	1	
	Intensity	18					1			
	Loudness	18								
	Formants (1-6)	108	29	6	1		43	14	4	
	MFCC	216	29	9			49	11	2	
EB	Left-Right mov.	30	6	2	1		9	2	1	
	Up-down mov.	30	1				5	2		
	Openness rate	30	7	1			9	2		
	Blinking	12	4	1			5	2	1	
	Looking Durations	24	12	5	1		13	7	4	
HB	Yaw mov.	60	7	4			13	7	4	1
	Roll mov.	60	5	1			15	9	6	2
	Pitch mov.	60	6	1			11	7	5	
	Mov. count totals	4	1				3			
	Total #features	814	147	53	9	0	241	106	51	14

*q#: is the level of intersection relaxation using q-relaxed method

the ability to generalise for modelling depression detection. Since we have 3 datasets and 1 for the combined, the q values in our investigation ranges from 0 (full intersection) to 3 (full union). The number of intersecting features from intersect and union modality-level aggregation from each dataset is detailed in Table 8.

Modality-level aggregation was performed in two approaches: intersection and union. Since the selected features from these two approaches differ in their aim and generalisation ability, we explore dataset-level aggregation in both of them. This is not to confuse the reader with the q-relaxed intersection in the dataset-level aggregation. For the q-relaxed intersection on the intersect features, when q is set to 0 (full intersection), the resulted feature set became empty, which is expected given the strictness of the intersect approaches from both modality-level and this level. Setting q value to 1 (i.e. agreement of all sets except 1), nine features remains, that include features from F0 (average, minimum and minimum Δ), HNR (minimum and range), shimmer (maximum $\Delta\Delta$), formants (2nd formants minimum), left-right eye movement (minimum), and eye gaze looking direction (maximum duration of looking down). With the increase of q value, more relaxation is introduced and therefore more features are included. Yet, the number of features being selected from each feature group indicates the strength of the feature group in distinguishing depression.

On the other hand, the q-relaxed intersection of the union features, a value of 0 set to q produced 14 features. These features introduced more feature groups and eliminated some others compared to $q1$ in the intersect features. These features include minimum, variance and standard deviation of F0 and both its derivatives, HNR minimum second derivative, average yaw and roll head movement and average speed of roll head movement. Since union features are already lenient, it can be noticed that the number of features is larger than the intersect features. Nonetheless, from the number of features selected in each group, it can be concluded that the strongest feature groups in characterising depression are F0, HNR, formants, MFCC, eye gaze direction, and head movement.

TABLE 9: Classification Results for each level of q-relaxed Intersection

	All-F	Intersect Features				Union Features			
Dataset	(814)	q3 (147)	q2 (53)	q1 (9)	q0 (0)	q3 (241)	q2 (106)	q1 (51)	q0 (14)
BD	58.3	71.7	80.0	78.3	-	75.0	75.0	73.3	51.7
AVEC	71.9	81.3	81.3	78.1	-	75.0	78.1	75.0	78.1
Pitt	79.0	81.6	81.6	81.6	-	84.2	76.3	76.3	68.4
All Datasets	74.6	76.2†	75.4	74.6	-	76.2†	75.4†	75.4†	61.5
Average	71.0	77.7	79.6	78.2		77.6	76.2	75.0	64.9

†: A fine (0.2) steps grid search was used to find the best parameters

6.4 Modelling Dataset-Level Features

We model depression severity detection using the refined selected features from the dataset-level aggregation. This is done using all level of q-relaxed method to assess the best level of relaxation in generalising and modelling depression (see Table 9).

Given that a value of 3 set to q-relaxed intersection means full union, the features selected in this relaxed level outperform using all features in all three datasets in both intersect and union features, and therefore, prove to have a generalisation ability. However, finding the optimal hyperplane using these sets of features were found with a fine search on the combined dataset, which could indicate a small margin of separation using these features. The raw values of the datasets were kept without further normalisation before combining the datasets, yet an optimal hyperplane was reached when using all features. On the other hand, a value of 0 for q-relaxed intersection results in a very small feature set from the union features (14 features) and an empty set for the intersection features. The results of $q0$ on union features did not generalise compared to when using all features set, except for AVEC dataset.

The q-relaxed intersection on different relaxation levels on the intersect features show improvement from using all features in all datasets, regardless of the reduction of the feature set size. For example, $q1$ on intersect features results in a feature set of 9 features, yet the classification results were marginally equal to $q2$ with 53 features. The same applies to the relaxed intersection on union features, except for the Pitt dataset with $q2$ and $q1$ levels. It can be concluded that even when the intersect features are strict, the model has better generalisation power to detect depression severity than is the case when using union features and all features. Considering the reduction in feature set size with the relaxation levels in the intersect features, $q1$ has the most satisfactory results.

7 Conclusion

In an effort to create an interpretative model of depression severity detection, we utilised feature selection method to identify the feature set that has not only a strong discrimination power in this task, but also has a generalisability to other depression datasets. For this purpose, we proposed a feature selection framework to select the most distinctive features using diverse feature selection methods from different families (i.e. streaming, structured and flat feature selection methods). We generalise the aggregated features from the framework from and on different datasets. The main findings are:

Feature Selection Methods and Framework: The proposed framework proves to be robust to randomness in selecting the features from different feature selection methods. We used stability measures to weigh the features selected from each method before aggregating the final selected features. Some feature selecting methods had a false high stability measure, where the framework was steady on selecting the features and limited the contribution of these feature selection methods. Structured data feature selection, in particular, network and tree-structured, were suitable for our feature space, which resulted in high stability and contributions from these methods.

Interpretation: After several levels of aggregations, features that have a high distinguishing ability for depression severity were identified. Subject speech behaviour and interaction prove to be the strongest in depression detection, however, not all dataset collection procedures included such interaction (i.e. AVEC). Therefore, generalisation of the approach on different datasets excluded the SB modality. The strongest speech prosody feature groups were F0, HNR, formants, and MFCC, while for eye activity modality they were left-right eye movement and gaze direction, and for head modality, it was yaw head movement.

Classification & Generalisation: The effectiveness of the selected features from different aggregation levels was investigated through modelling depression and generalising in different depression datasets. A feature set of 9 was able to generalise and outperform the modelling when using all feature space. Therefore, it could be concluded that the proposed framework has the ability to generalise on different and unseen datasets.

A limitation of this current work is the relatively small datasets, which did not allow a validation subset to reduce the contamination of seen data. However, to remedy and reduce the effect of this issue, several measures have been taken into consideration. First, feature selection methods were performed on 50 ensembles of several random splits of the training data (20% holdout). Second, these ensembles were run twice, where the intersecting features between the runs are selected. Third, the selected features from one dataset were used for generalisation on other unseen datasets. Even though future work could aim at investigating the effectiveness of the selected features on new depression datasets or new samples, acquiring such datasets is difficult, especially given the confidentiality of such datasets.

Nonetheless, improvement of the framework could be performed by identifying a threshold of feature selection methods for inclusion/exclusion. Moreover, refinement in the algorithm could include the rank/score of features as weight in the final feature selection. For simplicity, the current framework only uses binary weight (0: not selected, 1: selected).

Acknowledgement

This research was funded in part by the US National Institutes of Health grants MH51435, MH65376, and MH096951, and in part by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP190101294).

References

[1] T. Vos, C. Allen, M. Arora, R. M. Barber, Z. A. Bhutta, A. Brown, A. Carter, D. C. Casey, F. J. Charlson, A. Z. Chen *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015," *The Lancet*, vol. 388, no. 10053, pp. 1545–1602, 2016.

[2] A. Pampouchidou, P. Simos, K. Marias, F. Meriaudeau, F. Yang, M. Pediaditis, and M. Tsiknakis, "Automatic assessment of depression based on visual cues: A systematic review," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[3] 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 Communication*, vol. 71, pp. 10 – 49, 2015.

[4] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 578–584, Oct 2018.

[5] Y. Suhara, Y. Xu, and A. S. Pentland, "Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks," in *Proceedings of the 26th International Conference on World Wide Web*, ser. WWW '17. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017, pp. 715–724.

[6] T. Alhanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-September, pp. 1716–1720, 2018.

[7] M. A. H. M. S'adan, A. Pampouchidou, and F. Meriaudeau, "Deep learning techniques for depression assessment," in *2018 International Conference on Intelligent and Advanced System (ICIAS)*, Aug 2018, pp. 1–5.

[8] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding," *IEEE Transactions on Multimedia*, pp. 1–1, 2018.

[9] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1432–1441, July 2015.

[10] A. Mendiratta, F. Scibelli, A. M. Esposito, V. Capuano, L. Likforman-Sulem, M. N. Maldonato, A. Vinciarelli, and A. Esposito, *Automatic Detection of Depressive States from Speech*. Cham: Springer International Publishing, 2018, pp. 301–314.

[11] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of Biomedical Informatics*, vol. 83, pp. 103 – 111, 2018.

[12] S. Song, L. Shen, and M. Valstar, "Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018, pp. 158–165.

[13] K. Anis, H. Zakia, D. Mohamed, and C. Jeffrey, "Detecting depression severity by interpretable representations of motion dynamics," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018, pp. 739–745.

[14] K. Yu, X. Wu, W. Ding, and J. Pei, "Towards scalable and accurate online feature selection for big data," in *2014 IEEE International Conference on Data Mining*, Dec 2014, pp. 660–669.

[15] ——, "Scalable and accurate online feature selection for big data," *ACM Trans. Knowl. Discov. Data*, vol. 11, no. 2, pp. 16:1–16:39, Dec. 2016.

[16] X. Wu, K. Yu, H. Wang, and W. Ding, "Online streaming feature selection," *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, pp. 1159–1166, 2010.

[17] X. Wu, K. Yu, W. Ding, H. Wang, and X. Zhu, "Online feature selection with streaming features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1178–1192, 2013.

[18] J. Zhou, D. Foster, R. Stine, and L. Ungar, "Streamwise feature selection," *Journal of Machine Learning Research*, vol. 7, pp. 1861–1885, 2006.

[19] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine Learning*, vol. 65, no. 1, pp. 31–78, Oct 2006.

[20] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 171–234, Mar. 2010.

[21] R. Diaz-Uriarte and S. Alvarez de Andres, "Variable selection from random forests: application to gene expression data," *arXiv e-prints*, pp. q–bio/0 503 025, Mar. 2005.

[22] H. Deng and G. Runger, "Feature selection via regularized trees," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, June 2012, pp. 1–8.

[23] ——, "Gene selection with guided regularized random forest," *Pattern Recognition*, vol. 46, no. 12, pp. 3483 – 3489, 2013.

[24] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Grouped variable importance with random forests and application to multiple functional data analysis," *Computational Statistics & Data Analysis*, vol. 90, pp. 15 – 35, 2015.

[25] Z. Xu, G. Huang, K. Q. Weinberger, and A. X. Zheng, "Gradient boosted feature selection," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: ACM, 2014, pp. 522–531.

[26] G. Roffo, S. Melzi, and M. Cristani, "Infinite feature selection," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 4202–4210.

[27] G. Roffo, S. Melzi, U. Castellani, and A. Vinciarelli, "Infinite latent feature selection: A probabilistic latent graph-based ranking approach," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 1407–1415, 2017.

[28] G. Roffo and S. Melzi, "Features selection via eigenvector centrality," *Proceedings of New Frontiers in Mining Complex Patterns (NFMCP 2016)(Oct 2016)*, 2016.

[29] Q. Shen and A. Chouchoulas, "A modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems," *Engineering Applications of Artificial Intelligence*, vol. 13, no. 3, pp. 263 – 278, 2000.

[30] A. Janusz and D. Ślezak, "Random probes in computation and assessment of approximate reducts," in *Rough Sets and Intelligent Systems Paradigms*, M. Kryszkiewicz, C. Cornelis, D. Ciucci, J. Medina-Moreno, H. Motoda, and Z. W. Raś, Eds. Cham: Springer International Publishing, 2014, pp. 53–64.

[31] S. Zhao, E. C. C. Tsang, and D. Chen, "The model of fuzzy variable precision rough sets," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 2, pp. 451–467, April 2009.

[32] J. C. Davis, *Statistics and Data Analysis in Geology*, 2nd ed. New York, NY, USA: John Wiley & Sons, Inc., 1990.

[33] H. Liu and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," *Proceedings of the International Conference on Tools with Artificial Intelligence*, pp. 388–391, 1995.

[34] M. A. Hall and L. A. Smith, "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper." in *FLAIRS conference*, vol. 1999, 1999, pp. 235–239.

[35] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. New York, NY, USA: Wiley-Interscience, 2000.

[36] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine Learning*, vol. 53, no. 1, pp. 23–69, Oct 2003.

[37] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 1151–1157.

[38] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[39] H. H. Yang and J. Moody, "Data visualization and feature selection: New algorithms for nongaussian data," in *Advances in Neural Information Processing Systems*, 2000, pp. 687–693.

[40] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, pp. 1531–1555, 2004.

[41] P. Meyer, C. Schretter, and G. Bontempi, "Information-theoretic feature selection in microarray data using variable complementarity," *IEEE Journal on Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 261–274, 2008.

[42] R. Tibshirani, "The lasso method for variable selection in the cox model," *Statistics in medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[43] M. Y. Park and T. Hastie, "L1-regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 4, pp. 659–677, 2007.

[44] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[45] W. J. Fu, "Penalized regressions: The bridge versus the lasso," *Journal of Computational and Graphical Statistics*, vol. 7, no. 3, pp. 397–416, 1998.

[46] J. Yang and V. Honavar, *Feature Subset Selection Using a Genetic Algorithm*. Boston, MA: Springer US, 1998, pp. 117–136.

[47] M. Kursa and W. Rudnicki, "Feature selection with the boruta package," *Journal of Statistical Software*, vol. 36, no. 11, pp. 1–13, 2010.

[48] J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan, "Kernel feature selection via conditional covariance minimization," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 6949–6958.

[49] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, Jan 2002.

[50] A. Rakotomamonjy, "Variable selection using svm-based criteria," *Journal of Machine Learning Research*, vol. 3, pp. 1357–1370, 2003.

[51] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, "Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 478–490, Oct 2018.

[52] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 142–150, April 2013.

[53] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '13. New York, NY, USA: ACM, 2013, pp. 3–10.

[54] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber *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," *Biological psychiatry*, vol. 54, no. 5, pp. 573–583, 2003.

[55] J. C. Fournier, R. J. DeRubeis, S. D. Hollon, S. Dimidjian, J. D. Amsterdam, R. C. Shelton, and J. Fawcett, "Antidepressant drug effects and depression severity," *JAMA: The Journal of the American Medical Association*, vol. 303, no. 1, pp. 47–53, 2010.

[56] A. T. Beck, R. A. Steer, R. Ball, and W. Ranieri, "Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients," *Journal of personality assessment*, vol. 67, no. 3, pp. 588–597, Dec 1996.

[57] Depression Scores Conversion. (Online) Inventory of Depressive Symptomatology (IDS) and Quick Inventory of Depressive Symptomatology (QIDS).

[58] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "From joyous to clinically depressed: Mood detection using spontaneous speech," *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, FLAIRS-25*, pp. 141–146, 2012.

[59] E. A. Stepanov, S. Lathuilière, S. A. Chowdhury, A. Ghosh, R. Vieriu, N. Sebe, and G. Riccardi, "Depression severity estimation from multiple modalities," in *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, Sep. 2018, pp. 1–6.

[60] S. Alghowinem, R. Goecke, J. Epps, M. Wagner, and J. Cohn, "Cross-cultural depression recognition from vocal biomarkers," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-September-2016, pp. 1943–1947, 2016.

[61] N. Carvalho, E. Laurent, N. Noiret, G. Chopard, E. Haffen, D. Bennabi, and P. Vandel, "Eye movement in unipolar and bipolar depression: A systematic review of the literature," *Frontiers in Psychology*, vol. 6, p. 1809, 2015.

[62] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Eye movement analysis for depression detection," in *2013 IEEE International Conference on Image Processing*, Sep. 2013, pp. 4220–4224.

[63] S. Alghowinem, R. Goecke, J. Cohn, M. Wagner, G. Parker, and M. Breakspear, "Cross-cultural detection of depression from nonverbal behaviour," *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*, 2015.

[64] J. T. Fiquer, R. A. Moreno, J. Z. Canales, A. Cavalcanti, and C. Gorenstein, "Is nonverbal behavior in patients and interviewers relevant to the assessment of depression and its recovery? a study with dutch and brazilian patients," *Psychiatry Research*, vol. 250, pp. 59 – 64, 2017.

[65] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear, "Head pose and movement analysis as an indicator of depression," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Sep. 2013, pp. 283–288.

[66] H. Wang and C. Leng, "A note on adaptive group lasso," *Computational Statistics & Data Analysis*, vol. 52, no. 12, pp. 5277 – 5286, 2008.

[67] L. Jaulin, "Robust set-membership state estimation; application to underwater robotics," *Automatica*, vol. 45, no. 1, pp. 202 – 206, 2009.

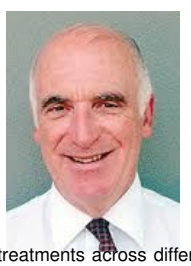
[68] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Characterising depressed speech for classification," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2534–2538, 2013.

Sharifa Alghowinem currently holds a postdoctoral fellow position at Massachusetts Institute of Technology. She worked as a lecturer at University of Canberra in 2011, research associate at Australian National University and hold a research and teaching position at Prince Sultan University. She received her PhD at the Australian National University, Computer Science Research School in 2015. She received her MSc in Software Engineering at University of Canberra in 2010, and her BSc in Computer Applications at King Saud University in 2004. Her research interests include speech processing, computer vision, affective computing, and machine learning.

Tom Gedeon received the BSc (Hons) and PhD degrees from the University of Western Australia. He is currently chair professor of Computer Science at the Australian National University, Canberra, Australia, and leads the Human-Centred Computing Group at the Research School of Computer Science. His research interests are in bio-inspired computing and in human centred computing. He is a former president of the Asia-Pacific Neural Network Assembly and a former President of the Computing Research and

Roland Goecke Roland Goecke is Professor of Affective Computing, Head of the Vision and Sensing Group, and Director of the Human-Centred Technology Research Centre at the University of Canberra. He received his Masters degree in Computer Science from the University of Rostock, Germany, in 1998 and his Ph.D. in Computer Science from the Australian National University, Canberra, Australia, in 2004. 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 are in affective computing, pattern recognition, computer vision, human-computer interaction and multimodal signal processing.

Jeffrey F. Cohn is a professor of psychology and psychiatry at the University of Pittsburgh and an adjunct professor at the Robotics Institute, Carnegie Mellon University. He has led interdisciplinary and inter-institutional efforts to develop advanced methods of automatic analysis and synthesis of facial and vocal expression and applied them to research in human emotion,interpersonal processes, social development,and psychopathology. He co-chaired the IEEE International Conference on Automatic Face and Gesture Recognition (FG2015 and FG2008), the International Conference on Multimodal Interfaces in 2014 (ICMI2014), and the International Conference on Affective Computing and Intelligent Interaction(ACII2009). His research is supported in part by the US National Institutes of Health and the US National Science Foundation. He is an associate member of the IEEE.



Gordon Parker is Scientia Professor of Psychiatry at 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 at UNSW and Director of Psychiatry at 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, Parker was elected as a Fellow of the Academy of the Social Sciences in Australia, and in 2010 was recipient of an Officer of Order of Australia Award for distinguished service to psychiatry.