Looking At The Body: Automatic Analysis of Body Gestures and Self-Adaptors in Psychological Distress

Weizhe Lin, *Member, IEEE,* Indigo Orton, Qingbiao Li, Gabriela Pavarini, and Marwa Mahmoud, *Member, IEEE*

Abstract—Psychological distress is a significant and growing issue in society. In particular, depression and anxiety are leading causes of disability that often go undetected or late-diagnosed. Automatic detection, assessment, and analysis of behavioural markers of psychological distress can help improve identification and support prevention and early intervention efforts. Compared to modalities such as face, head, and vocal, research investigating the use of the body modality for these tasks is relatively sparse, which is partly due to the limited available datasets and difficulty in automatically extracting useful body features. To enable our research, we have collected and analyzed a new dataset containing full body videos for interviews and self-reported distress labels. We propose a novel approach to automatically detect self-adaptors and fidgeting, a subset of self-adaptors that has been shown to correlate with psychological distress. We perform analysis on statistical body gestures and fidgeting features to explore how distress levels affect behaviors. We then propose a multi-modal approach that combines different feature representations using Multi-modal Deep Denoising Auto-Encoders and Improved Fisher Vector Encoding. We demonstrate that our proposed model, combining audio-visual features with detected fidgeting behavioral cues, can successfully predict depression and anxiety in the dataset.

Index Terms—Self-adaptors, fidgeting, psychological distress, digital phenotyping, behavioural sensing

1 Introduction

Psychological distress and mental disorders are significant threats to global health [1].[1] According to the World Health Organization (WHO), an estimated 450 million people around the world suffer from neuropsychiatric conditions [3], with depression and anxiety being the most common mental disorders [4]. Despite existing strategies for the treatment of distress, such as depression, it is estimated that nearly two-thirds of people suffering distress have never received help from a health professional [5]. Early detection of distress is consistently noted as a key factor in treatment and positive outcomes [6], [7]. Early detection requires an ongoing assessment to identify distress when it begins. Self-evidently, ongoing assessment at scale is prohibitive when performed manually. As such, automatic detection of signs of psychological distress or specific mental disorders is an active area of research [8], [9].

Mental health assessments are largely based on self-reports and health workers' subjective observations. Automated detection of behavioural markers of distress, for instance based on video data, can also help add greater objectivity and complement these assessments, supporting health professionals in decision-making. Such automated analysis might be particularly helpful for early clinicians or lay health workers without extensive training in psychiatry. Automated video analysis is also relevant in remote monitoring and assessment of individuals at risk [10]. Since the COVID-19 pandemic, telepsychiatry has seen an exponential growth, and automated analysis can help address long-standing concerns about the ability of health professionals to pick up subtle behavioural signs remotely e.g. during a video call [11]. In addition to supporting clinicians, visualizations of the analysis can be fed-back to participants and used for structured self-reflection during a therapeutic session [12].

Currently, the most effective automated distress detection approaches utilize multi-modal machine learning. These modalities include facial, head, eye, linguistic (textual), vocal, and body.

There are significant challenges to body modality research, particularly within automatic distress detection, including the lack of relevant data, the inability to share much of the data, and the difficulty in gathering such data. Specifically, the combination of full-body data (either sensor-based or video-based) with psychological distress labels is rare. Compounding this rarity is the private and sensitive nature of the data, which means such datasets are rarely shared publicly.

Body expressions, and especially self-adaptors, have been shown to be correlated with human affect, depression and psychological distress [13], [14], [15], [16], [17]. Self-adaptors are self-comforting gestures, including any kind of touching on other parts of the body, either dynamically or statically [18], [19]. Fidgeting, a subset of self-adaptors, is the act of moving about restlessly, playing with one's fingers,

- *W. Lin is with the Department of Engineering, Cambridge. E-mail: wl356@cam.ac.uk*
- *I. Orton, Q. Li and M. Mahmoud are with the Department of Computer Science and Technology, Cambridge. Email: {indigo.orton@cl., ql295@, mmam3@}cam.ac.uk*
- *G. Pavarini is with the Department of Psychiatry, Oxford. Email: gabriela.pavarini@psych.ox.ac.uk*

1. This work is an extension of the work in [2], originally published in the proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG) 2020

hair, or personal objects in a way that is not peripheral or nonessential to ongoing tasks or events [20]. Patients with depression often engage in self-adaptors [21]. Fidgeting is a sign of attention-deficit and hyperactivity disorder, also exhibited by individuals with autism [22]. With manually annotated data, Scherer *et al.* [23] reported a longer average duration of self-adaptors as well as fidgeting for distressed participants, while hand-tapping was reported to correlate with depression and anxiety [19].

More recent advances in the state-of-the-art for pose estimation [24] enable accurate pose data on a broader set of datasets and thus open the door for new approaches for body expression analysis and broader incorporation of body features in multi-modal systems.

In this paper, we propose to use a hierarchical model to automatically detect self-adaptors as well as fidgeting, which has been shown to be predictive of psychological distress. We analyzed body gestures and self adaptors in a dataset of video recordings that we collected, concentrating on symptoms of depression and anxiety because these are the most common mental disorders [4]. We then present two methods to explore the body modality (especially fidgeting): First with a statistical linearity analysis with traditional linear regression, and second with a deep-learning-based pipeline. In the second method, a Multi-modal Deep Denoising Auto-Encoder (multi-DDAE) is utilized for encoding per-frame features. Improved Fisher Vector encoding [25] is then used to generate per-sample representation. Finally, we demonstrate that these features are discriminative in psychological distress detection. We refer to psychological distress as a state of emotional suffering, including symptoms of anxiety, depression and psychological stress [26], varying from normal adjustment issues to diagnosable mental health conditions [27].

The contributions of this paper can be summarized as follows:

1) We introduce a new audio-visual dataset containing recordings of non-clinical interviews along with distress labels from established psychological evaluation questionnaires.

2) We propose a hierarchical model for automatic detection of self-adaptors (including fidgeting) from visual data and evaluate our approach on a publicly available fidgeting dataset with manual labels.

3) We present a statistical analysis of a set of statistical body gesture features as well as specific fidgeting features extracted from the body modality data and explore how distress levels affect participants' behavior in our dataset.

4) As proof of concept, we implement a multi-modal feature fusion framework to perform distress classification and demonstrate the importance of self-adaptors features, specifically fidgeting, in predicting symptoms of depression and anxiety.

2 Related Work

In this section, we focus on related work on automatic detection of signs of psychological distress, including studies that focus on separate modalities and multi-modal fusion frameworks.

2.1 Facial and head modality

Facial Action Coding System (FACS) [28] has long been used to taxonomize human facial movements by their appearance on the face, which yields the concept of Facial Action Units (AUs). For example, the Audio/Visual Emotion Challenge (thereafter AVEC) used AUs features as a basic descriptor for its psychological distress detection tasks.

A big body of literature has been developed to analyze facial expressions and the head modalities in the context of depression and psychological distress. For example, Yang *et al.* [29] proposed a "Histogram of Displacement Range (HDR)", which is a measurement of the amount of facial landmark movements to predict depression. Joshi *et al.* [30] presented a categorization analysis framework which consists of "bag of facial dynamics" and "histogram of head movements". Dibeklioğlu *et al.* [31] [32] feature-engineered dynamic representation (e.g., velocity, acceleration, and standard deviation of motion) for facial landmark movement and head motion and used them in a multi-modal system to detect depression in a dataset of clinical interviews.

Psychomotor retardation refers to a slowing-down of thought and a reduction of physical movements in an individual. Sobin *et al.* [33] demonstrated the correlation between psychomotor retardation and depression. Syed *et al.* [34] handcrafted descriptors using craniofacial movements in order to capture the psychomotor retardation, and then made predictions of depression.

Some other features such as lower emotional expressivity [35], eye lid movement [34], reduced gaze activity [36] [37], and averted gaze [35] have been also used as predictive features of depression.

2.2 Audio modality

Acoustic features of speech can be predictive of distress irrespective of the speech content [38], [39]. For example, Ozdas *et al.* [38] assessed the risk of suicide by detecting the fluctuations in the fundamental frequency of people's speech. Dibeklioğlu *et al.* [31] explored the use of vocal prosody for depression detection. Similarly, Syed *et al.* [34] investigated the use of turbulence in speech patterns.

Besides, in AVEC challenges, low-level descriptors of voice signals, such as Mel-frequency Cepstral Coefficients (MFCCs), are provided, leading to many multi-modal methods incorporating these acoustic features for distress and mental illness detection [29], [40].

2.3 Body modality

A few previous studies attempted to include the body modality in their models to predict psychological distress, mostly by extracting generic features from the video recordings related to the body. For example, Joshi *et al.* [30] computed Histogram of Gradients (HOGs) and Histogram of Optical Flow (HOFs) around the generic Space-Time Interest Points (STIPs) extracted from the videos, and then generated a "Bag of Body Dynamics" feature that was used for depression classification. Some of the multi-modal work presented in the AVEC challenges [40], [41], [42], [43] utilize the low-level descriptors of visual signals (such as latent

CNN layer activation of ResNet [44] and VGGNet [45]) to predict on psychological distress.

More recent works also investigate the specific movement of body parts. In the past few years, the skeletal models, either using RGB such as OpenPose [24] or RGBD such as Microsoft Kinect SDK skeleton tracker[2], have gained popularity for action recognition tasks and were used to generate more dedicated and explainable features (rather than those global generic descriptors such as STIPs) using the motions of specific body parts by feature engineering [18], [46]. For example, Jaiswal *et al.* [46] extracted head movements using Kinect and performed multi-modal classification with other audiovisual features to predict Attention Deficit Hyperactivity Disorder (ADHD) and Autism Spectrum Disorder (ASD). Though promising, the related work using such skeletal models on detecting psychological distress is still sparse.

In terms of automatic detection of self-adaptors, the only previous work that attempted to detect fidgeting behavior was presented by Mahmoud *et al.* [18]. They developed a multi-modal framework for automatic detection of descriptors of rhythmic body movement by extracting Speeded-Up Robust Features (SURFs) interest points around Microsoft Kinect pose points and then detected rhythmic behaviors from analyzing the trajectories of the interest points. However, there are two limitations in their proposed automated system when applied to distress detection: 1) The dataset they used was based on acted data, so the behavior detected is not natural. For example, in more real interview scenarios, participants do not always fidget with a rhythmic pattern. 2) The trajectory data was noisy, and their method could not sufficiently handle the complexity of the detected body signal. As such, they were only able to achieve 59% recognition on their acted dataset.

2.4 Multi-modal Learning

Since psychological distress is expressed through all modalities, many of the state-of-the-art models that predict signs of psychological distress proposed multi-modal approaches [29], [40], [40], [41], [42], [43], [47], [48], combining low-level features extracted from the face, speech, and text, which are usually the features publicly available for the datasets. By only working with extracted features, most of these works focused on exploiting the given features, instead of analyzing the behavioral cues (e.g., specific gestures) of psychological distress. For example, the winner of AVEC 2019 [42] proposed multi-layer attention fusion frameworks, but they did not explore the psychological basis of their models' decisions due to the lack of access to the raw data.

3 Dataset

In this section, we describe the data collection, experimental design, and general characteristics of our collected dataset. This dataset is designed to enable investigation of the body modality for use in automatic detection of distress.

2. https://developer.microsoft.com/en-us/windows/kinect/

3.1 Overview and design

Participants were recruited through the University of Cambridge email lists, student social media groups, and paper fliers posted around the town. We aimed to balance the sample with regards to distress levels, such that the database includes participants at the two distinct ends of the distress spectrum. To identify participants with high versus low levels of distress, we conducted an online screening with a total 106 people who signed up for the study. Participants completed standardized measures of depression (PHQ-8 [49], [50]) and anxiety (GAD-7 [51]), as well as demographics. In the selection, we balanced the participants according to the public norm shown in Table 2 (e.g., for depression, above 6.63 is marked as high, otherwise low). Given potential gender differences in nonverbal communication [52], we also balanced the final sample with regards to gender within each distress group[3]. From the initial screening, 35 were invited to the face to face session, including 18 with high distress and 17 with low distress.

The participants completed the same measures of depression and anxiety immediately before the interview. This was meant to provide an assessment of distress closer in time to the interview and to increase the psychological salience of this information during the interview.

We adopted a data collection methodology inspired by the DAIC dataset collection method [53], which consists of a human interviewer asking a series of open-ended conversational questions to elicit naturalistic behavior. The interviews were performed by a computer science researcher based on peer-support interview questions collected from the university support services. To achieve the conversational interview dynamic the interviewer asks general questions regarding the participant's life and further encourages the participant to elaborate. For example, the interviewer would ask "can you tell me about one time in your life you were particularly happy?" and then ask some follow up questions regarding the example the participant provided. The interviewer was blind to the distress level of participants during the interview.

To keep behaviors naturalistic, participants were not aware of the main goal of the study, which is an automatic analysis of behavioral cues. Instead, they were told that the experiment aimed at building models that can help in mental well-being. This ensured that their behavior would be as natural as possible. All participants got debriefed of the main aim of the data collection at the end of the session. Participants were not informed of the results of their questionnaires, and all of them were handed a small booklet with the list of peer support and mental well-being services provided by the university. It is worth mentioning that the interviewer was blind to whether participants were from high or low distress groups in order not to affect their behavior. They were also instructed to limit their body and facial expressions throughout the interview and keep their sitting posture constant through all the interviews in

3. Non-binary/other was given as an option in the registration form. A number of people registered with this option. However, none of those people met the distress level criteria and were thus not selected for an interview.

Label	Range	Mean	Covariance with Depression
Distress			
Depression	0–19	7.43	-
Anxiety	0–19	7.00	86.15%
Perceived stress	1–30	18.17	84.00%
Somatic symptoms	1–27	9.06	74.16%
Personality			
Extraversion	3–31	16.37	-30.49%
Agreeableness	12–34	25.67	-42.21%
Openness	7–39	27.29	4.29%
Neuroticism	1–31	16.86	80.00%
Conscientiousness	10–36	21.46	-46.41%
Demographic			
Gender	18 M & 17 F		9.47%
Age	18–52	25.40	-11.09%

TABLE 1
Descriptive statistics regarding the questionnaire and demographic results within the dataset. This table demonstrates there are no confounding correlations with the depression label.

order to avoid any changes in participants' behavior due to mimicry effect [54].

The dataset is labeled with participant responses to self-evaluation questionnaires right before the interview for assessing distress and personality traits, as well as demographic labels such as gender. The distress questionnaires include PHQ-8 for depression, GAD-7 for anxiety, SSS-8 [55] for somatic symptoms, and PSS [56] for perceived stress. Personality traits are measured using the Big Five Inventory [57]. In sum, each participant provided responses to 5 questionnaires, in which PHQ-8 and GAD-7 were measured twice, both at registration and before the face-to-face session.

As a result, the dataset includes videos of fully natural non-acted expressions, including facial expressions, body motion, gestures, and speech.

3.2 Preliminary Analysis

We collected videos of 35 interviewed participants with a total video duration of 07:50:08 (hours:minutes:seconds). Descriptive statistics regarding the questionnaire and demographic results within the dataset are provided in Table 1. Covariance is presented as normalized covariance values, also known as the correlation coefficient.

Confounding Correlations

We assessed confounding correlations based on the depression label, as much of the related work focuses on depression. While the distress measures, anxiety, perceived stress, and somatic stress, were found to be strongly correlated with depression, the personality measures have below 50% covariance with the exception of neuroticism, which is a trait characterized by negative emotionality, with an 80% covariance. The demographic measures, gender, and age were negligibly correlated, with 9.47% and -11.09% covariance, respectively. Finally, the interview duration was found to be not correlated with any questionnaire result (less than 25% covariance with all labels). Thus, we can be confident that there are no confounding correlations with personality scores or demographics.

Published Norms

A comparison of the mean values for distress and personality measures between our dataset and the published norms is presented in Table 2. While there are differences, the measures are generally in line with the published norms. The dataset has a substantially higher mean perceived stress score, but only slightly higher mean scores for anxiety and depression. Depression, extraversion, and neuroticism measures are particularly close to their published norms. While the dataset mean for agreeableness and openness are substantially higher than the published norms (over 10% over the technical range for those measures).

Label	Mean	Norm	Source
Distress			
Depression	7.43	6.63	Ory et al. [58]
Anxiety	7.00	5.57	Spitzer et al. [51]
Perceived stress	18.17	12.76	Cohen et al. [56]
Somatic symptoms	9.06	12.92	Gierk et al. [55]
Personality			
Extraversion	16.37	16.36	Srivastava et al. [59]
Agreeableness	25.67	18.64	Srivastava et al. [59]
Openness	27.29	19.61	Srivastava et al. [59]
Neuroticism	16.86	16.08	Srivastava et al. [59]
Conscientiousness	21.46	18.14	Srivastava et al. [59]

TABLE 2
Comparison of the mean questionnaire values within our dataset to the published norms. This shows that the population distribution, with regards to these distress and personality measures, is generally in line with the broader population.

3.3 Remarks

Participants completed the PHQ-8 and GAD-7 questionnaires twice: during registration and with the interview process. These questionnaires are temporal; specifically, they relate to the participant's mental state in the past two weeks. Given this, some difference between registration and interview results was expected.

With the exception of a small number of outliers, participants were generally consistent in self-evaluation between registration and interview. PHQ-8 responses had a mean difference of 0.89, while GAD-7 responses had a mean difference of 0.63. As a result, we took the most recent response to self-evaluation questionnaires as the label for each participant's video recording.

The dataset features, labels and/or videos will be shared with the research community on a case-by-case basis by request.

4 METHOD

We used our collected dataset to study body gestures and self-adaptors. In this section, we demonstrate two different methods to analyze the body modality within the context of psychological distress. As a first step, we extract the most common audio-visual features. Then we describe a set of generic statistical body features that we extract to analyze general body gesture movement. To look specifically for self-adaptors, we then present an automatic approach to extract self-adaptors and fidgeting behavior in our dataset. We then perform a feature-based statistical analysis on the extracted body features - both generic and fidgeting

features to understand what features are generally correlated with distress classification. Lastly, we move on to propose a multi-modal approach to demonstrate further the effectiveness of body modality, where we incorporate and analyze the co-occurrence of multiple modalities to make predictions.

4.1 Audiovisual Feature Extraction

4.1.1 Visual Features

For each video, we used state-of-the-art tools, OpenPose [24] and OpenFace 2.2 [60], to extract body pose features, facial Action Units (AUs), and gaze directions.

However, OpenPose and OpenFace do not take into account the consistency of the keypoints across time, causing the keypoints to usually fluctuate highly in many parts, introducing noise to the real continuous face and body motion. Besides, there are some frames where OpenPose or OpenFace fail to extract all pose points or gaze features, respectively. To overcome these problems, we infer the missing data via Cubic Spline Interpolation across the whole sequence. We then smooth the data using a Savitzky-Golay filter [61] (window length is 11 and the order of the polynomial is 3).

4.1.2 Audio Features

Speaker diarization involves partitioning an audio stream into homogeneous segments according to the speaker's identity. In order to distinguish the speech of the interviewer and the participant, we use the open-source Speaker-Diarization project [62] which utilizes an Unbounded Interleaved-State Recurrent Neural Network (UIS-RNN) [63], to extract speaker identities with respect to the time axis. We then conduct a manual check to assign correct diarization labels to the participant and the interviewer. We also use pyAudioAnalysis [64] to extract MFCCs.

4.2 Generic Body Features

To explore the body modality, we extract and analyze the set of generic statistical features that describe the body movements.

4.2.1 Feature Extraction

Two kinds of statistical features are computed and extracted: global features and localized features. In the global features, we care about the overall statistics of motion, while in the localized features (features that are within specific body parts, such as head, hands, and legs), we are interested in the statistics of the motion within the body parts, which we refer to as "localization". Our notation is summarized in Table 3.

We define a "gesture" as a period of sustained movement within a body localization. For example, waving hands is a gesture within "`Hn` (hand)" localization, and shaking legs continuously will register a gesture in "`L` (Legs)" localization.

To detect gestures within a localization, we scan the video using a sliding window method.

First, the per-frame absolute movement (L^2 distance) is calculated for each pose point. The value is then averaged by the number of pose points in the localization. Formally,

$$F_t = \frac{1}{|P|} \sum_{p \in P} ||P_{p,t} - P_{p,t-1}||_2 \quad (1)$$

where $P_{p,t}$ is the position vector of pose point p at time t, and F_t is the averaged per-frame movement across all points. P are the collection of pose points in this localization.

Second, a sliding window is applied such that a small number of frames do not have a disproportionate effect on the detection. This process can be expressed by:

$$W_i = \frac{1}{l} \sum_{t=i \times l}^{t<i \times (l+1)} F_t \quad (2)$$

where W_i is the windowed average at window index i, l is the length of the window, and F_t is the average movement at frame t, from Equation 1. We experimentally chose $l = 10$, i.e. a second of movement is represented by 3 windows.

Third, the window moves until an average movement above a threshold is found, which is considered the beginning of the gesture. The gesture continues until $n = 3$ consecutive windows (30 frames, approximately 1 sec) are found below the movement threshold, which is thus considered the end of the gesture.

Localization	Abbr.
Overall	`O`
Hands	`Hn`
Head	`He`
Legs	`L`

	Feature	Abbr.
Overall	Average **F**rame **M**ovement	`FM`
	Proportion of total **M**ovement occurring during a **G**esture	`GM`
	Average **G**esture **S**urprise	`GS`
	Average **G**esture movement standard **D**eviation	`GD`
	Number of **G**estures	`GN`
Localized	Average **L**ength of **G**esture	`GL`
	Average per-frame **G**esture movement	`GA`
	Total movement in **G**estures	`GT`
	Average **G**esture **S**urprise	`GS`
	Number of **G**estures	`GN`

TABLE 3
Feature notation Abbrs. of `BodyGesture`.

Table 3 lists the set of body features we extract. Below we explain how we define each of these features for the overall body. Similarly, the localized features can be calculated for every localization/body part.

- **Average frame movement** - the per-frame average movement (moving distance) of every pose point of the body. This is the only feature that is not based on detected gestures.
- **Total movement** - the sum of the absolute movements of all points within the sliding window.
- **Proportion of total movement occurring during a gesture** - the proportion of total movement that occurred while a gesture is happening (within some localizations).

- **Average gesture surprise** - defined as "fraction of frames with no gesture happening" ÷ "number of gestures".
  For example, if two gestures occurred within a sample such that 80% of the sample duration had no gesture occurring, the average gesture surprise would be $\frac{80\%}{2} = 40\%$. Whereas, if there were 100 gestures, the average surprise is 0.8%, even though both samples had the same proportion without any gesture occurring. This matches the intuition that each gesture within 100 evenly spaced gestures would be unsurprising as they were regularly occurring, whereas the 2 evenly spaced gestures would be surprising because nothing was happening in between.
- **Average gesture movement standard deviation** - the standard deviation of per-frame movement within a gesture is averaged across all detected gestures. This is intended to indicate the consistency of movement intensity through a gesture.
- **Number of gestures** - the total number of detected gestures across all tracked localizations.

4.2.2 Feature Processing

All the movement data is extracted from smoothed OpenPose data described in Section 4.1.1. As described in Table 3, there are 5 Overall features (`O-FM/GM/GS/GD/GN`) and 15 localized features (`Hn/He/L` each followed by `-GL/GA/GT/GS/GN`). All these body gesture features are concatenated (thereafter marked as `BodyGesture`, which has a feature vector of length 20 for each participant) and all features are normalized such that the length of the sample does not affect the results.

Sum-based features (e.g., gesture length, gesture count, total movement, etc.) are normalized against the total number of frames in the sample. Gesture average features, such as gesture surprise, are again normalized against the total number of gestures.

4.3 Self-adaptors and Fidgeting Features

In addition to the generic body features, we were interested in analyzing the self-adaptors and fidgeting behavior. In this section, we present our fidgeting detection system in three subsections. We start by exploring the self-adaptors/fidgeting encoding and the overall hierarchical design. Then we show the methods of building the two essential detectors of our hierarchical model in the following two subsections. For each detector, we demonstrate the detector's design, and then present the labeling strategy which provides reliable labels for training and evaluation. In order to validate the effectiveness of our automated fidgeting detection approach before moving onto distress classification, we evaluate our model thoroughly both on an acted dataset and on our newly collected dataset of natural expressions.

4.3.1 Overall Design and Encoding

Given the lack of broad agreement on the definition of fidgeting so far, we utilize a two-step hierarchical model to identify fidgeting. The overall hierarchical design of our self-adaptor/fidgeting detector is presented in Fig. 1 and the encoding scheme is shown in Table 4. We first identify self-adaptors, which we define as low-level location events (e.g. `H2H`, `H2F` as in Table 4). Secondly, action events (i.e. DYNAMIC, STATIC) of hand/leg are classified by the *DYNAMIC/STATIC Classifier*. Fidgeting is then defined as a combination of low-level self-adaptors and action events. Specifically, we define three types of fidgeting: cross hand fidgeting, single-hand fidgeting, and leg/feet fidgeting.

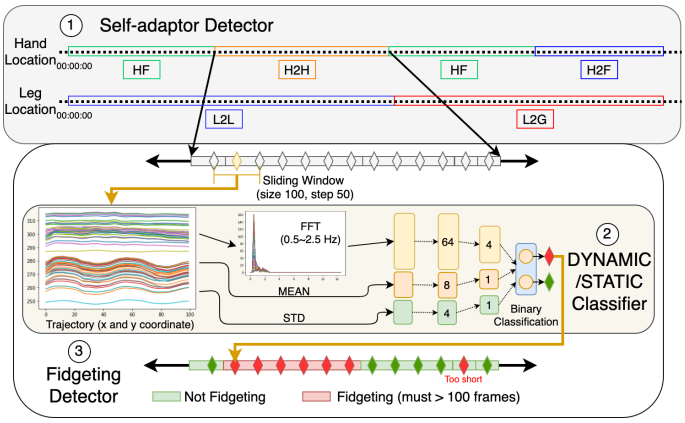

Fig. 1. Hierarchical self-adaptor/fidgeting detection workflow. (1) First, detect hand/leg location; (2) Classify motion within each sliding window using *DYNAMIC/STATIC Classifier*; (3) Finally, combine location and motion to give high-level fidgeting event. The figure shows the detection of `H2H` (Hand to hand) fidget. The same principle applies to other fidgets.

Self-adaptors	**Description**
`H2H`	**H**and to **H**and
`H2A`	**H**and to **A**rm
`H2L`	**H**and to **L**eg
`H2F`	**H**and to **F**ace
`HF`	**H**and **F**ree (when not belong to any of above)
`L2G`	Both **L**egs on **G**round
`L2L`	**L**eg on the other **L**eg (crossed legs)
Action Events	**Description**
DYNAMIC	Moving obviously
STATIC	No obvious movement is observed

Fidgeting Type	**Combination**
`CHF` (**C**ross **H**and **F**idgeting)	`H2H` + DYNAMIC
`SHF` (**S**ingle **H**and **F**idgeting)	{`H2A`, `H2L`, `H2F`, `H2F`} + DYNAMIC
`SHF-L` (to **L**eg only)	`H2L` + DYNAMIC
`SHF-F` (to **F**ace only)	`H2F` + DYNAMIC
`SHF-A` (to **A**rm only)	`H2A` + DYNAMIC
`LFF` (**L**eg/**F**eet **F**idgeting)	{`L2G`, `L2L`} + DYNAMIC

TABLE 4
Self-adaptor and fidgeting encoding book

4.3.2 Self-adaptor Detector

4.3.2.1 Design: Each body location is represented using a bounding box. Self-adaptors are defined as overlapping bounding boxes. We represent the hand and face using the smallest rectangular box bounding all corresponding hand or face keypoints. The long sides of bounding boxes for the forearms, upper arms, lower legs, and upper legs are aligned with the connection between two joints from

OpenPose, while the width is a free parameter tuned for the best detection performance.

First, H2H self-adaptor events are detected (i.e., when the two hands' bounding boxes overlap). Then all other hand-based self-adaptor events are detected, for all segments of the video not containing H2H segments.

All self-adaptors, except for H2F, must be longer than 100 frames (around 4 seconds with the frame rate of 26). This reduces the noise from detected self-adaptor events.

4.3.2.2 Labeling and Evaluation: In order to validate our self-adaptor detector, we manually labeled 4 participants' videos, a total duration of 59 minutes. The inter-labeler agreement between 3 annotators was checked using Krippendorff's alpha. Each frame was labeled with one of the self-adaptor codes from Table 4. Within these videos, participants perform different self-adaptors and each event has a minimum total duration of 5 minutes, with the exception of H2F which is less frequent in contrast to others.

As shown in Table 5, the Krippendorff's alpha agreement for left-hand location is 0.823, for right-hand location is 0.888 and for leg location is 1.00. This suggests good agreement between the annotators and, thus, the reliability of the labels. The results show that our design that utilizes OpenPose and the interactions between bounding boxes is able to detect self-adaptor with excellent overall precision, and especially for the H2H, H2F, L2L and L2G events, the detector reached a very high accuracy. Note that, 'NA' in Table 5 means that there is no corresponding gestures in the evaluation set of 4 labelled participants.

Hand Self-adaptors (left/right)

	Precision	Recall	F1 Score
H2H	1.00/1.00	0.99/0.99	1.00/1.00
H2A	1.00/NA	0.64/NA	0.79/NA
H2L	0.96/0.88	0.86/0.82	0.91/0.85
H2F	NA/1.00	NA/1.00	NA/1.00
HF	0.63/0.83	0.99/0.98	0.77/0.90
Alpha Score:	0.823/0.888		

Leg Location

	Precision	Recall	F1 Score
L2L	1.00	1.00	1.00
L2G	1.00	1.00	1.00
Alpha Score:	1.000		

TABLE 5
Self-adaptor Detection Evaluation

4.3.3 Fidgeting Detector

4.3.3.1 Design: As shown in Fig. 1, the *DYNAMIC/STATIC Classifier* operates on extracted **optical flow** from a sliding window across the video (size 100 frames, step 50 frames). To classify the action (DYNAMIC/STATIC), hand movements (especially fingers) and leg movements require optical flow to obtain smooth trajectories, given OpenPose estimations become unreliable when hands intersect or are occluded. We thus initialize the optical flow with the OpenPose estimations at the beginning of each slice.

We choose Fast Fourier Transform (FFT), standard deviation (STD), and mean values (MEAN) of point trajectories as our input features (in this case, number of trajectories is 2 $\times$ number of keypoints as we have 2-D data for each keypoint). For fidgeting, we are more interested in the cyclic motion with a frequency ranging from 0.5Hz to 2.5Hz [18]. Therefore, we extracted the spectrum data within the range $[0.5, 2.5]$ Hz. As we analyze slices of length 100, the dimension of FFT spectrum data that is within $[0.5, 2.5]$ Hz is always fixed at $41\times$ number of trajectories. An FFT feature of length 41 is obtained by averaging over FFT values of trajectories that have the same frequency. As for the STD and MEAN features, we simply calculate along the time axis and give a vector with a length of the number of trajectories for each feature.

4.3.3.2 Labeling and Evaluation: To train and evaluate the *DYNAMIC/STATIC Classifiers*, accurate labeling is required. Three classifiers are required to cover the three categories of detected self-adaptors: {H2H}, {H2A, H2L, H2F, HF}, and {L2G, L2L}.

We labelled DYNAMIC/STATIC on each of the three categories. We randomly sampled and labeled approximately 30% of slices for each category in every video.

Two researchers labeled the data independently. As shown in Table 6, we first manually dropped the slices with a wrong category label (e.g. a slice is detected as H2H while it's in fact not). The number of slices that have a correct category label is shown as "Correct". Secondly, we labeled DYNAMIC/STATIC and dropped the slices that lack a consensus between two researchers. The number of slices with an agreement is shown as "Agreed". The high percentage of both "Correct" and "Agreed" suggests the good performance of our self-adaptor detection and also the high reliability of action labels.

Category	Total	Correct	Agreed
BOTH: H2H	3962	3922 (99%)	3793 (96%)
LEFT:{H2A, H2L, H2F, HF}	1614	1566 (97%)	1539 (96%)
RIGHT:{H2A, H2L, H2F, HF}	1620	1588 (98%)	1563 (96%)
{L2G, L2L}	6536	6536 (100%)	6196 (95%)

TABLE 6
Hand/Leg action labelling overview. The values in the table are the number of slices generated by the sliding window.

Having reliable slice labels, we then partitioned participants into 5 folds and performed slice-level cross-validation. For evaluation, we calculated accuracy, F1 score, and their respective standard deviations.

Category	Acc.	Acc. Std.	F1	F1 Std.
BOTH: H2H	0.833	0.019	0.834	0.019
LEFT:{H2A, H2L, H2F, HF}	0.884	0.025	0.884	0.026
RIGHT:{H2A, H2L, H2F, HF}	0.895	0.026	0.894	0.026
{L2G, L2L}	0.875	0.022	0.871	0.021

TABLE 7
DYNAMIC/STATIC Classifier evaluation (LEFT means left hand, RIGHT means right hand, BOTH means both hands)

As shown in Table 7, the detector achieved generally high accuracy and F1 score with low standard deviations. Though the hand actions are difficult even for researchers to label, the detector can successfully classify more than 80% of slices.

4.4 Feature encoding

This section describes how we encoded low-level frame-level features described in Sec 4.1 and 4.3 in preparation for the final prediction step. The generic statistical

Feature Group	Dimension	Description
BodyGesture	20×1	Body Gesture Statistical Features
Fidget	$9 \times N$	Fidget feature & Speaking array
Fidget_pure	$8 \times N$	Fidget feature only
Gaze	$8 \times N$	Gaze direction
AUs	$35 \times N$	Action Units
MFCCs	$13 \times N$	Acoustic features

TABLE 8
Feature Groups. N is number of frames in each recording of participants.

BodyGesture will not need to be encoded since it represents global statistical features rather than time-series features.

4.4.1 Fidgeting features processing

Having extracted low-level features from each frame, we combine them to form high-level descriptors of fidgeting behavior (CHF, SHF, and LFF as shown in Table 4). The Fidget_pure feature group is formed by {HCF, SHF-L(left hand), SHF-L(right hand), SHF-A(left hand), SHF-A(right hand), SHF-F(left hand), SHF-F(right hand), LFF}. The Fidget_pure group is combined with a participant speaking feature array to form the full fidget feature group, enabling us to investigate whether fidgeting and speaking co-occurrence is relevant. This participant speaking feature array indicates whether the participant is speaking during a frame. This is calculated using the previously described diarization data.

After all the feature extraction, we have several feature groups shown in Table 8.

4.4.2 Per-frame representation

In order to capture more useful feature representations and reduce the dimensionality, and inspired by our previous work [65], different modalities are combined using a Multimodal Deep Denoising Auto-Encoder (multi-DDAE). As shown in Fig. 2, each modality is encoded through a dense layer and then all are concatenated to yield the last shared dense layer which provides the representation we use. The shared layer is then inversely decoded to generate each modality. We optimized the hyper-parameters of the auto-encoder via several experiments so that the dimensions of hidden layers are $\{0.5d, 0.25d, 0.5d\}$ where d represents the input dimension of each node, and the noise applied at the input is 0.1 Gaussian noise. The training optimization target is the joint Mean Square Error (MSE) of the MSEs of the feature group at each node (later we fixed the loss weights to be 0.35 for the fidget feature group while 0.1 for others, as we are more interested in fidgeting in our experiments).

4.4.3 Whole video representation

Due to varying lengths of the videos, it's necessary to unify the dimensionality of the per-video representation. Though Fisher Vector was originally proposed to aggregate visual features [25], it has become popular in social signal processing such as bipolar disorder [66] and depression recognition [67]. Inspired by these applications, we apply a Gaussian Mixture Model to cluster similar per-frame representations and then use an Improved Fisher Vector encoding to obtain a fixed-length representation. As a result, the feature is transformed from num_frames × feature_dim to 2 × GMM_Kernel_num × feature_dim.

Feature Set	F1-Score
O-FM	34.43%
BodyGesture	66.81%
Searched BodyGesture	82.70%
Fidget_pure	49.60%
Searched [BodyGesture, Fidget_pure]	**83.38%**

TABLE 9
Results of linear regression threshold classification on body gesture statistical features and fidget features. [A, B] represents a concatenation of feature vector A and B.

4.5 Classification of signs of distress

We apply a Random Forest to select important features from the per-video representation. The selected features are used by the classifier. We experiment with two classifiers: 1) a logistic regression-based classifier (LR) using a binary threshold of 0.5; 2) a Multi-Layer Perception (MLP) with two softmax outputs for binary classification (number of layers are shown in Fig. 2).

As the available samples are limited and the useful features vary across individual differences, label smoothing [68] is applied to the MLP model in order to further boost the performance. More formally:

$$L\_new = L \times (1 - s) + \frac{s}{n} \tag{3}$$

where L is the one-hot label at softmax outputs, s is the smoothing parameter, and n is the number of classification classes. For example, when smoothing is 0.2, the one-hot label $\{0, 1\}$ will become $\{0.1, 0.9\}$, which lowers the confidence on training samples but reduces overfitting.

5 Statistical Analysis of Body Gesture

To better understand the effect of different body-related features, before moving to deep multimodal learning, we deploy a simple linear regression model to perform statistical analysis on the body gesture features (BodyGesture from Sec. 4.2) and fidgeting features (Fidget_pure from Sec. 4.4). The aim of this section is to shed some light on the effect of different movements of every part of the body and its correlation with depression.

5.1 Experimental Setup

Fidgeting features from Fidget_pure is processed by averaging along the time axis ($9 \times N$ to 9×1) to match the dimension of other features in BodyGesture (20×1). Reporting notation is defined as "[localization]-[feature type][linear polarity]". Localization and feature type token mappings are provided in Table 3. Polarity is defined below:

- "+/¬": A greater value (e.g. more activity) contributing to a positive/negative classification
- "/": A near-zero coefficient in linear model.
- "?": The polarity is observed inconsistent in different folds of cross-validation.

With the linear model, we perform 3-fold cross-validation on depression labels, which is more reliable than normal train-valid-test split for our small dataset. Cross-validation also provides more confidence about the polarity of each feature, as only the features that show consistent polarity across all folds will be marked. All results are calculated as the mean of 3-fold cross-validation results. All experiments and cross-validation are participant-independent.

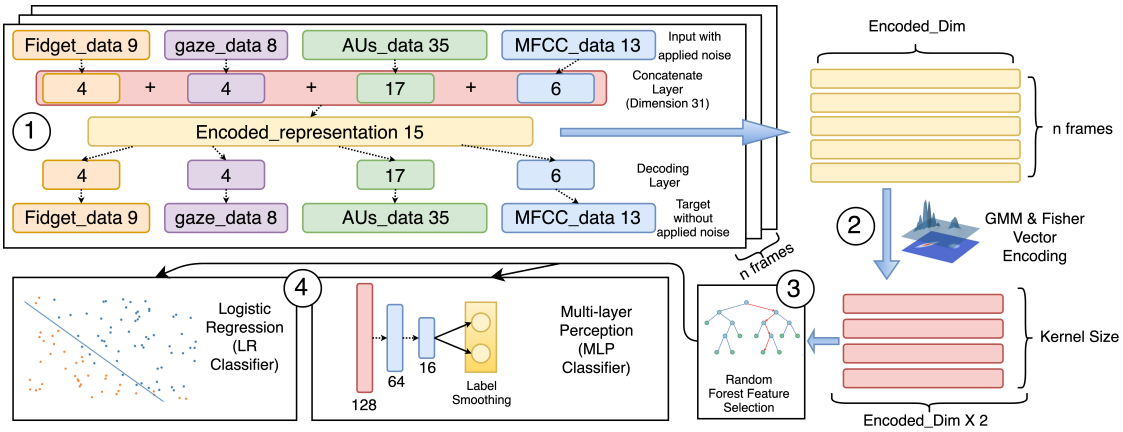

Fig. 2. Multi-modal fusion & classification pipeline. The dashed arrow represents a fully connected neural network between dense layers. Pose estimation, gaze, Action Units, and MFCC data are extracted from videos. Fidget features are computed using the method described in Section 4. (1) All features are fed into a Multi-modal Deep Denoising Auto-Encoder (multi-DDAE) to generate a compact per-frame encoded representation. (2) These per-frame features are then compressed into a whole video representation using a Gaussian Mixture Model (GMM) and Fisher Vector combination. (3) Random Forest feature selection is performed. (4) Finally, a classifier predicts a given label. We experiment with two classifiers, a logistic regression classifier and a Multi-layer Perception.

5.2 Results and Discussion

As shown in Table 9, with only the global movement (`O-FM`), the F1 score is only 34.43%. This means that measuring the quantity of global motion in the body is not a good indicator of depression. While when combining all body gesture statistical features, the classifier achieves 66.81% F1 score.

Note that all body gesture statistical features include a large set of features representing statistics of different body parts as well as global body motion, as explained in Sec. 4.2. In order to filter out this large feature set, we performed an exhaustive feature search to obtain the combination of features that gives the best performance, represented in Table 9 as "Searched `BodyGesture`". It reaches a good F1 score at 82.70%.

As shown in Table 9, when we combine specific fidgeting features (`Fidget_pure`) with `BodyGesture`, and perform feature search on the concatenated feature, the F1-score reaches the best at 83.38%. The resulted best feature combination includes: {`O-FM?, O-GM+, O-GN?, Hn-GN?, Hn-GS¬, He-GL+, He-GN+, He-GT+, He-GA+, He-GS+, L-GL+, L-GN+, L-GA+, SHF-L(Right)+, SHF-A(Right)+, SHF-F(Right)+, SHF-F(Left)+`}. Looking deeply into this list of features we could infer some interesting insights into the overall body movements in our dataset, which we explain below.

For example, the `O-GM+` token suggests that more movement within gestures relative to all other movement is indicative of depression, and especially, total movement within head gestures (`He-GT+`) is positively correlated with depression. The localized features suggest that the length of gestures in the head and legs (`He-GL+, L-GL+`) is correlated with depression. It's clear that gesture statistics in hands (`Hn-⋆`) are generally not interesting in prediction, while the classifier pays more attention to head and leg motions. However, `Hn-GS¬` suggests that more regular (thus less surprising) hand gestures (e.g. constant fidgeting) show a positive contribution to depression.

We can also conclude that a higher quantity of right hand fidgeting on the leg, arm, and face (`SHF-⋆(Right)+`) have a positive contribution to the higher depression level, and left hand fidgeting on the face (`SHF-F(Left)+`) is also positively correlated with high depression level. The difference in left and right arise from the fact that most participants are right-handed and thus their left hands exhibit less useful motions that are predictive of depression. This conclusion is not surprising, as, in our observations, people perform hand to hand fidgeting regardless of their depression label. Combining the results from above, we can conclude that, in our dataset, more regular hand gestures and more fidgeting on the leg, arm, and face are indicative of depression. Depressed participants also have exhibit frequent motions in the head and leg region.

6 Evaluation of Multimodal Deep Learning

In this section, we evaluate and demonstrate the validity and potential of fidgeting features as complementing modality with other features to predict the signs of psychological distress.

First, we present some baseline distress classification results on our dataset. Next, we present results for our full multi-modal classifier pipeline, where we investigate the effects of hyper-parameters on the performance given the small size of our dataset. Finally, we apply our automatic fidgeting detection approach to a publicly available dataset [18] to demonstrate its accuracy and generalisability beyond our dataset.

As in Sec. 5, all results are calculated as the mean of 3-fold cross-validation results. All experiments and cross-validation are participant-independent.

6.1 Baselines

As a baseline, we used Gaussian kernel Support Vector Machines (SVMs) classifiers applied on each individual feature group used in our multi-modal model (listed in Table 8). Unlike in Sec. 5, non-linearity can be considered in these baseline models. They are evaluated for a binary depression label and a binary anxiety label. These models provide a simple and common baseline for our dataset. For the baseline SVM, we use the mean value for each

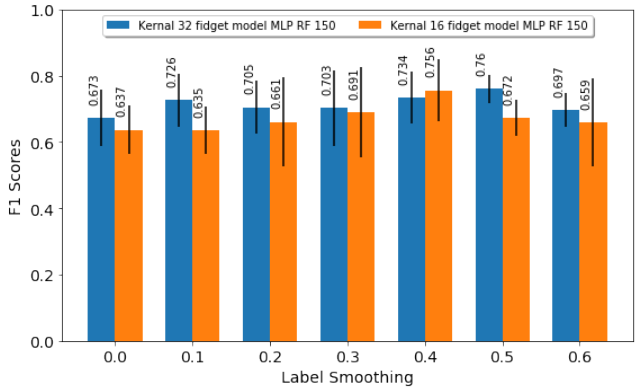

Fig. 3. Effects of label smoothing. In general, smoothing can boost performance. (error bar extends by the standard deviation in either side and best performance in **bold**)

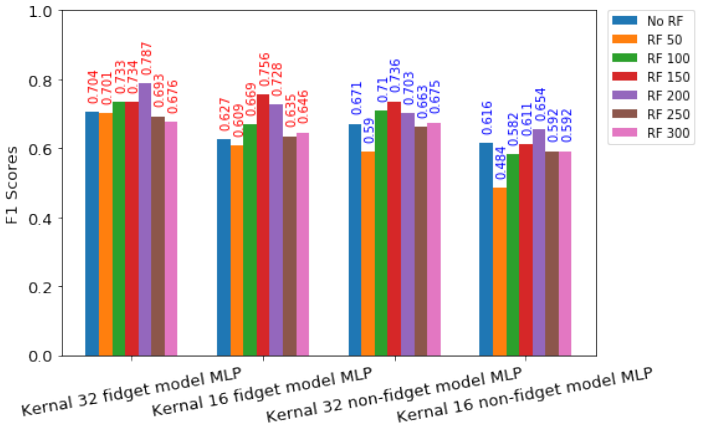

Fig. 4. Effects of hyper-parameters. Red denotes models incorporating fidget features and blue for non-fidget models. In general, models with fidget features perform better. (Error bars are not shown for better visualization; best performance of each model is in **bold**). RF+number denotes the number of features selected by Random Forest.

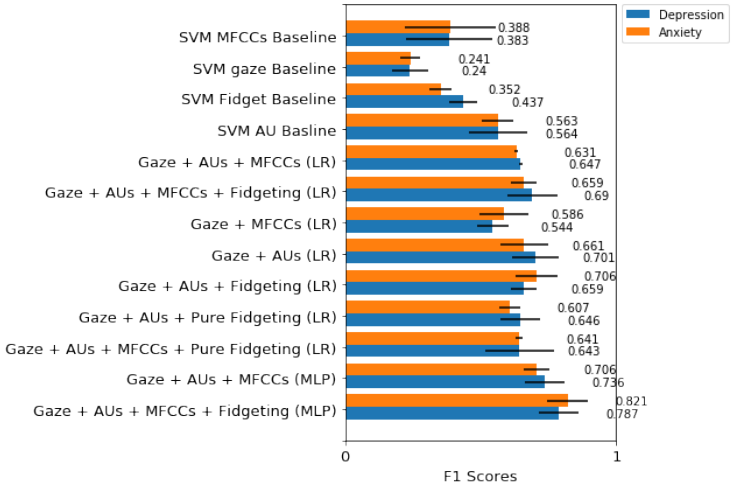

Fig. 5. Effects of feature groups and ablation analysis (error bars extend by the standard deviation in either side; best performance is in **bold**).

feature over the whole sample, thus providing a normalized representation with mean values of all the features. Results are presented in Fig. 5.

These baseline models demonstrate two points: first, the behaviors we are attempting to classify in our dataset are complex; and second, our fidgeting features by themselves are not trivially predictive of distress, but rather require learned representations.

6.2 Multi-modal distress classification

As presented in the previous baseline section, single modalities are not enough to capture the complexity of signs of psychological distress. Therefore we experiment with our proposed multi-modal classification framework. We encode different modalities through multi-DDAE and Improved Fisher Vector encoding (Sec. 4.4), and classify distress labels using either LR or MLP classifier after Random Forest feature selection (Sec. 4.5).

In Fig. 5, we present the best performance of different feature group combinations using our multi-modal fusion framework. We use a Random Forests (RF) for feature selection. As RFs take in labels to find the most discriminative features, this feature selection is only performed on the training set and selected features are then applied to the test set, which prevents label leaking.

6.2.1 Effects of some hyper-parameters

As shown in Fig. 3, when other hyperparamters are fixed, label smoothing makes great effects on classification performance. Fig. 3 presents the great effect of label smoothing on classification performance when other hyperparameters are fixed. Though some turbulences exist, the performance increases with higher label smoothing but starts to decrease when smoothing is too much. This is intuitively reasonable because when smoothing is above 0.5, there is less allowed space for model to learn features well. The results in Fig. 3 shows that label smoothing parameter at 0.4 generally provides good performance, and thus we fixed this value in all following experiments.

We test different numbers of features selected by RF (RF_num), and different GMM kernel sizes. Fig. 4 shows that the performance is generally worse when RF_num is low (< 100) as it results in insufficient information with most of the features unselected. However, when RF_num is high (≥ 250), redundant features bias the classifier, decreasing performance.

Using 32 GMM kernels achieves better performance than 16 kernels. We hypothesize that this improvement stems from incorporating more GMM components for clustering similar per-frame features. More kernels enable more clusters and thus more predictive features. However, when kernel size is above 32, the fitting score is large (in GMM lower is better) and therefore increasing beyond 32 will not further improve performance.

6.2.2 Effects of feature groups

From Fig. 5, it is clear that fidget features improve most configurations' performance, but performance decreases marginally without the participant speaking event (presented as "Pure Fidgeting" in figure). Therefore, we can conlude that the co-occurrence of speaking and fidgeting is beneficial for distress detection.

6.2.3 Ablation Analysis

Fig. 5 also demonstrates our ablation studies to help us analyze the important factors in distress classification. We

remove one or two feature groups from our framework and conduct the same experiments.

Without MFCCs features, the performance generally doesn't drop too much in depression and even increases in anxiety. This suggests that MFCCs are not very important in depression and even distractive in anxiety detection.

AUs have long been proved to be predictive of distress, and, as expected, we see a significant performance reduction when omitting them.

It is interesting to note that fidgeting, with the LR configuration, does not consistently improve performance, but in anxiety, it always boosts the classification performance. Therefore, we conclude that fidgeting is certainly important in anxiety, but is also predictive in depression when combined with other feature configuration.

6.3 Fidget detector cross-dataset validation

To further validate our automatic fidgeting detection approach, we evaluate it on a publicly available dataset from Mahmoud *et al.* [18] that has videos of fidgeting behavior along with manual fidgeting labels.

In this dataset, actors perform specific fidgets. While these fidgets are overemphasized compared to natural fidgets, their core movement is similar.

Segments of the video containing fidgeting are manually labeled in an action-exclusive manner. That is, the co-occurrence of fidgeting is not labeled. Given this, we measure the accuracy of our approach in two phases: first, we check that fidgeting, regardless of location, is detected during the periods of manually labeled fidgeting; and second, we calculate the recall for location-specific fidgeting. Precision would not make sense for location-specific fidgeting, because the detected location may also be fidgeting, while the ground truth only considers one location.

Detected fidgeting segments shorter than 100 frames are excluded to reduce noise. As shown in Table 10, the recall of

Step 1: Detect fidget only

Fidget	Precision	Recall	F1-Score	Support
0	0.51	0.49	0.50	29440
1	0.79	0.80	0.80	69517

Step 2: Detect specific fidgeting (evaluated with recall)

Fidget type	Recall	Support
leg	0.784	32430
hand to face	0.865	10594
hand to arm	0.787	12794
hand cross	0.768	13699

TABLE 10
Results of fidget detection on Mahmoud *et al.*'s dataset [18]. Support refers to the total number of samples.

the non-fidget label is around 50%, but this is due to the fact that the labels are generally assigned to a long continuous segment and do not accurately reflect the actions occurring per-frame. However, the recall of the fidget label is good, achieving 80%.

Our fidgeting detection approach outperforms the state-of-the-art presented by Mahmoud *et al.* [18] for each fidget type, achieving a recall above 75% for all fidgeting types.

7 Conclusion

In this paper, we conducted a pioneering research on automatic detection of body gesture descriptors as a subset of behavioural markers of psychological distress that can be used in integrated tools to complement experts' assessment and support health professionals. We introduced a novel audio-visual distress dataset comprising recorded interviews and distress labels based on psychological questionnaires, where we investigated the relationship between body gestures and psychological distress.

We then presented an automated self-adaptor and fidgeting detection approach to extract different fidgeting behaviors trained on real interview videos. Our approach outperformed the state-of-the-art method when evaluated on a manually-labeled publicly-available fidgeting dataset. It was also successful in detecting fidgeting behaviour in our newly collected dataset of natural expressions.

Statistical analysis with a large set of generic gesture features was carried out, providing interesting insights into the effect of different generic body movements and their correlation with depression levels.

We also presented a deep learning approach for psychological distress detection that doesn't require a feature search and utilizes the co-occurrence of different multimodal features. The system successfully detected depression and anxiety with around 80% F1-scores, and an ablation study has been carried out demonstrating the great value of fidgeting behavior descriptors in predicting signs of psychological distress.

8 Limitations and Future Work

Despite the limitation of the small dataset we used, our work demonstrate the importance of the fidgeting features as a complementary modality for classification and prediction of psychological distress.

In our multi-modal classification experiments, we treated all fidgeting features as a whole. For future work, it will be interesting to evaluate the importance of each fidget behavior (e.g., hand to arm fidget and hand to hand fidget). In our work, we only focused on depression and anxiety disorders. However, our automatic approach to detecting self-adaptors and fidgeting opens the door for more work to explore the presence of these non-verbal behaviors and measure them quantitatively in other psychological disorders. The code and pre-trained models for our fidgeting detection system are already available to the research community at Github[4], which enables further research in this field.

References

[1] T. Vos, A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas, F. Abd-Allah, R. S. Abdulkader, A. M. Abdulle, T. A. Abebo, S. F. Abera *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet*, vol. 390, no. 10100, pp. 1211–1259, 2017.

[2] W. Lin, I. Orton, M. Liu, and M. Mahmoud, "Automatic detection of self-adaptors for psychological distress," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2020.

4. https://github.com/LinWeizheDragon/AutoFidgetDetection

[3] World Health Organization (WHO), *The World Health Report 2001: Mental health : new understanding, new hope*. Geneva: World Health Organization(WHO), 2017.

[4] WHO, *Depression and other common mental disorders: global health estimates*. World Health Organization(WHO), 2001.

[5] Mental Health Foundation, "Home - mental health foundation," *Mental Health Foundation*, accessed 2020/07/04. [Online]. Available: https://www.mentalhealth.org.uk/

[6] M. G. Craske and B. G. Zucker, "Prevention of anxiety disorders: A model for intervention," *Applied and Preventive Psychology*, vol. 10, no. 3, pp. 155–175, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0962184901800123

[7] C. G. Davey and P. D. McGorry, "Early intervention for depression in young people: a blind spot in mental health care," *The Lancet Psychiatry*, vol. 6, no. 3, pp. 267–272, 2019.

[8] Y. Lee, R.-M. Ragguett, R. B. Mansur, J. J. Boutilier, J. D. Rosenblat, A. Trevizol, E. Brietzke, K. Lin, Z. Pan, M. Subramaniapillai, T. C. Chan, D. Fus, C. Park, N. Musial, H. Zuckerman, V. C.-H. Chen, R. Ho, C. Rong, and R. S. McIntyre, "Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review," *Journal of Affective Disorders*, vol. 241, pp. 519–532, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165032718304853

[9] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," *Current Opinion in Behavioral Sciences*, vol. 18, pp. 43–49, 2017, big data in the behavioural sciences. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352154617300384

[10] A. Pampouchidou, M. Pediaditis, E. Kazantzaki, S. Sfakianakis, I. Apostolaki, K. Argyraki, D. Manousos, F. Meriaudeau, K. Marias, F. Yang *et al.*, "Automated facial video-based recognition of depression and anxiety symptom severity: cross-corpus validation," *Machine Vision and Applications*, vol. 31, no. 4, pp. 1–19, 2020.

[11] S. Dave, S. Abraham, R. Ramkisson, S. Matheiken, A. S. Pillai, H. Reza, J. Bamrah, and D. K. Tracy, "Digital psychiatry and covid19: The big bang effect for the nhs?" *BJPsych Bulletin*, pp. 1–11, 2020.

[12] K. Huckvale, S. Venkatesh, and H. Christensen, "Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–11, 2019.

[13] B. De Gelder, "Why bodies? Twelve reasons for including bodily expressions in affective neuroscience," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3475–3484, 2009.

[14] M. Neff, N. Toothman, R. Bowmani, J. E. F. Tree, and M. A. Walker, "Don't scratch! Self-adaptors reflect emotional stability," in *International Workshop on Intelligent Virtual Agents (IVA)*. Springer, 2011, pp. 398–411.

[15] M. Mahmoud and P. Robinson, "Towards automatic analysis of gestures and body expressions in depression." in *PervasiveHealth*, 2016, pp. 276–277.

[16] K. Chui, C.-Y. Lee, K. Yeh, and P.-C. Chao, "Semantic processing of self-adaptors, emblems, and iconic gestures: An erp study," *Journal of Neurolinguistics*, vol. 47, pp. 105–122, 2018.

[17] S. Chan, M. Khader, J. Ang, J. Chin, and W. Chai, "To behave like a liar: Nonverbal cues to deception in an asian sample," *Journal of Police and Criminal Psychology*, vol. 31, no. 3, pp. 165–172, 2016.

[18] M. Mahmoud, L.-P. Morency, and P. Robinson, "Automatic multimodal descriptors of rhythmic body movement," in *International Conference on Multimodal Interaction (ICMI)*. ACM, 2013, pp. 429–436.

[19] L. A. Fairbanks, M. T. McGuire, and C. J. Harris, "Nonverbal interaction of patients and therapists during psychiatric interviews." *Journal of Abnormal Psychology*, vol. 91, no. 2, p. 109, 1982.

[20] A. Mehrabian and S. L. Friedman, "An analysis of fidgeting and associated individual differences," *Journal of Personality*, vol. 54, no. 2, pp. 406–429, 1986.

[21] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Nonverbal Communication, Interaction, and Gesture*, vol. 1, no. 1, p. 57–106, 1969.

[22] J. M. Froiland and M. L. Davison, "Home literacy, television viewing, fidgeting and adhd in young children," *Educational Psychology*, vol. 36, no. 8, pp. 1337–1353, 2016.

[23] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–8.

[24] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[25] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision (ECCV)*. Springer, 2010, pp. 143–156.

[26] J. L. Burnette, L. E. Knouse, D. T. Vavra, E. O'Boyle, and M. A. Brooks, "Growth mindsets and psychological distress: A meta-analysis," *Clinical Psychology Review*, vol. 77, p. 101816, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0272735820300040

[27] P. Supportive, P. C. E. Board *et al.*, "Adjustment to cancer: anxiety and distress (pdq®)," *PDQ Cancer Information Summaries [Internet]*, 2015.

[28] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, 1978.

[29] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*. ACM, 2017, pp. 53–59.

[30] J. Joshi, R. Goecke, G. Parker, and M. Breakspear, "Can body expressions contribute to automatic depression analysis?" in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–7.

[31] H. Dibeklioğlu, Z. Hammal, Y. Yang, and J. F. Cohn, "Multimodal detection of depression in clinical interviews," in *International Conference on Multimodal Interaction (ICMI)*. ACM, 2015, pp. 307–310.

[32] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 525–536, 2017.

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

[34] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression severity prediction based on biomarkers of psychomotor retardation," in *Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*. ACM, 2017, pp. 37–43.

[35] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, L.-P. Morency *et al.*, "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing*, vol. 32, no. 10, pp. 648–658, 2014.

[36] S. Alghowinem, R. Goecke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear, "Cross-cultural detection of depression from nonverbal behaviour," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–8.

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

[38] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Analysis of fundamental frequency for near term suicidal risk assessment," in *International Conference on Systems, Man and Cybernetics (SMC)*, vol. 3. IEEE, 2000, pp. 1853–1858.

[39] N. Srimadhur and S. Lalitha, "An end-to-end model for detection and assessment of depression levels using speech," *Procedia Computer Science*, vol. 171, pp. 12–21, 2020.

[40] X. Xing, B. Cai, Y. Zhao, S. Li, Z. He, and W. Fan, "Multi-modality hierarchical recall based on gbdts for bipolar disorder classification," in *Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*. ACM, 2018, pp. 31–37.

[41] L. Yang, Y. Li, H. Chen, D. Jiang, M. C. Oveneke, and H. Sahli, "Bipolar disorder recognition with histogram features of arousal and body gestures," in *Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*. ACM, 2018, pp. 15–21.

[42] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," in *Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2019, pp. 81–88.

[43] S. Yin, C. Liang, H. Ding, and S. Wang, "A multi-modal hierarchical recurrent neural network for depression detection," in *Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2019, pp. 65–71.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.
[46] S. Jaiswal, M. F. Valstar, A. Gillott, and D. Daley, "Automatic detection of adhd and asd from expressive behaviour in rgbd data," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2017, pp. 762–769.
[47] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud *et al.*, "AVEC 2018: Bipolar disorder and cross-cultural affect recognition," in *Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*. ACM, 2018, pp. 3–13.
[48] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "AVEC 2017: Real-life depression, and affect recognition workshop and challenge," in *Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*. ACM, 2017, pp. 3–9.
[49] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, no. 1-3, pp. 163–173, Apr. 2009.
[50] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.
[51] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder," *Archives of Internal Medicine*, vol. 166, no. 10, pp. 1092–1097, May 2006.
[52] A. Mehrabian, *Nonverbal communication*. Transaction Publishers, 1972.
[53] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. R. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview Corpus of human and computer interviews." *International Conference on Language Resources and Evaluation (LREC)*, 2014.
[54] U. Hess, P. Philippot, and S. Blairy, "Mimicry: Facts and fiction," *The Social Context of Nonverbal Behavior*, pp. 213–241, 1999.
[55] B. Gierk, S. Kohlmann, K. Kroenke, L. Spangenberg, M. Zenger, E. Brähler, and B. Löwe, "The somatic symptom scale–8 (SSS-8)," *JAMA Internal Medicine*, vol. 174, no. 3, pp. 399–407, Mar. 2014.
[56] S. Cohen, T. Kamarck, and R. Mermelstein, "Perceived stress scale," *Measuring Stress: A Guide for Health and Social Scientists*, vol. 10, pp. 1–2, 1994.
[57] O. P. John, S. Srivastava *et al.*, "The big five trait taxonomy: history, measurement, and theoretical perspectives," *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.
[58] M. G. Ory, S. Ahn, L. Jiang, K. Lorig, P. Ritter, D. D. Laurent, N. Whitelaw, and M. L. Smith, "National study of chronic disease self-management: six-month outcome findings," *Journal of Aging and Health*, vol. 25, no. 7, pp. 1258–1274, Sep. 2013.
[59] S. Srivastava, O. P. John, S. D. Gosling, and J. Potter, "Development of personality in early and middle adulthood: Set like plaster or persistent change?" *Journal of Personality and Social Psychology*, vol. 84, no. 5, pp. 1041–1053, 2003.
[60] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2018, pp. 59–66.
[61] R. W. Schafer *et al.*, "What is a savitzky-golay filter," *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 111–117, 2011.
[62] L. Dong, "Speaker diarizationn," *Github repository*, accessed 2020/07/04. [Online]. Available: https://github.com/taylorlu/Speaker-Diarization
[63] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6301–6305.
[64] T. Giannakopoulos, "pyAudioAnalysis: an spen-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.
[65] Z. Zhang, W. Lin, M. S. Liu, and M. Mahmoud, "Multimodal deep learning framework for mental disorder recognition," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2020.
[66] Z. S. Syed, K. Sidorov, and D. Marshall, "Automated screening for bipolar disorder from audio/visual modalities," in *Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*. ACM, 2018, pp. 39–45.
[67] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 255–259.
[68] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 4696–4705.

Weizhe Lin is a PhD student in Information Engineering at the University of Cambridge. He studied Computer Science at Hong Kong University for one year, and then transferred to Cambridge for a 4-year BA/MEng course in Engineering. He was awarded a silver medal in "Future Scientist Award Program" by Chinese Academy of Sciences and Academy of Engineering. His research interest spans Natural Language Processing, Hyperspectral Image Processing, and Affective Computing.

Indigo Orton is a PhD student in Computer Science at the University of Cambridge. From Australia, Indigo completed his Bachelor in Information Technology at Deakin University in Melbourne, Australia. He then spent a number of years in industry building successful startups in the USA and Australia. Returning to academia, Indigo received his MPhil in Computer Science from Cambridge in 2019 and later that year commenced his PhD.

Qinbiao Li is a Ph.D. student in the Prorok Lab at the University of Cambridge. Before joining Cambridge, he completed an MRes degree in Medical Robotics at Imperial College London, and an MEng degree in Mechanical Engineering at the University of Edinburgh. His research interests span Robot Learning (IL and RL approaches), Graph Neural Networks (GNNs), and machine learning applications in the medical imaging area.

Dr Gabriela Pavarini is a Postdoctoral Research Fellow at the Department of Psychiatry at the University of Oxford. Her current research focuses on adolescent mental health, peer-led interventions and ethics of new technologies in psychiatry. She co-designs digital tools such as games and chatbots to facilitate youth engagement in the design and implementation of mental health care.

Dr Marwa Mahmoud is a Research Fellow of King's College and an Affiliated Lecturer at the Department of Computer Science and Technology, University of Cambridge. Her research focuses on computer vision and machine learning within the context of affective computing, behaviour analytics and human behaviour understanding. She is particularly interested in building inference models that tackle challenging real-world problems, usually characterised by data scarcity and noisy signals from multiple modalities. She applies her research in the areas of automotive applications, healthcare, and animal welfare.