

Explainable Depression Detection via Head Motion Patterns

Monika Gahalawat
University of Canberra
monika.gahalawat@canberra.edu.au

Raul Fernandez Rojas
University of Canberra
raul.fernandezrojas@canberra.edu.au

Tanaya Guha
University of Glasgow
tanaya.guha@glasgow.ac.uk

Ramanathan Subramanian
University of Canberra
ram.subramanian@canberra.edu.au

Roland Goecke
University of Canberra
roland.goecke@canberra.edu.au

ABSTRACT

While depression has been studied via multimodal non-verbal behavioural cues, head motion behaviour has not received much attention as a biomarker. This study demonstrates the utility of fundamental head-motion units, termed *kinemes*, for depression detection by adopting two distinct approaches, and employing distinctive features: (a) discovering kinemes from head motion data corresponding to both depressed patients and healthy controls, and (b) learning kineme patterns only from healthy controls, and computing statistics derived from reconstruction errors for both the patient and control classes. Employing machine learning methods, we evaluate depression classification performance on the *BlackDog* and *AVEC2013* datasets. Our findings indicate that: (1) head motion patterns are effective biomarkers for detecting depressive symptoms, and (2) explanatory kineme patterns consistent with prior findings can be observed for the two classes. Overall, we achieve peak F1 scores of 0.79 and 0.82, respectively, over *BlackDog* and *AVEC2013* for binary classification over episodic *thin-slices*, and a peak F1 of 0.72 over videos for *AVEC2013*.

KEYWORDS

Kinemes, Head-motion, Depression detection, Explainability

1 INTRODUCTION

Clinical depression, a prevalent mental health condition, is considered as one of the leading contributors to the global health-related burden [19, 26], affecting millions of people worldwide [31, 44]. As a mood disorder, it is characterised by a prolonged (> two weeks) feeling of sadness, worthlessness and hopelessness, a reduced interest and a loss of pleasure in normal daily life activities, sleep disturbances, tiredness and lack of energy. Depression can lead to suicide in extreme cases [18] and is often linked to comorbidities such as anxiety disorders, substance abuse disorders, hypertensive diseases, metabolic diseases, and diabetes [10, 41]. Although effective treatment options are available, diagnosing depression through self-report and clinical observations presents significant challenges due to the inherent subjectivity and biases involved.

Over the last decade, researchers from affective computing and psychology have focused on investigating objective measures that can aid clinicians in the initial diagnosis and monitoring of treatment progress of clinical depression [11, 33]. A key catalyst to this progress is the availability of relevant datasets, such as *AVEC2013* and subsequent challenges [43]. In recent years, research on depression detection employing affective computing approaches has increasingly focused on leveraging non-verbal behavioural cues such as facial expressions [9, 13], body gestures [23], eye gaze [4],

head movements [5] and verbal features [12, 22] extracted from multimedia data to develop distinctive features to classify individuals as depressed or healthy controls, or to estimate the severity of depression on a continuous scale.

In this study, we examine the utility of inherently interpretable head motion units, referred to as *kinemes* [27], for assessing depression. Initially, we utilise data from both healthy controls and depressed patients to discover a basis set of kinemes via the (*pitch*, *yaw*, and *roll*) head pose angular data obtained from short overlapping time-segments (termed two-class kineme discovery or 2CKD). Further, we employ these kinemes to generate features based on the frequency of occurrence of distinctive, class-characteristic kinemes. Subsequently, we discover kineme patterns solely from head pose data corresponding to healthy controls (Healthy control kineme discovery or HCKD), and use them to represent both healthy and depressed class segments. A set of statistical features are then computed from the reconstruction errors between the raw and learned head-motion segments corresponding to both the depressed and control classes (see Figure 1). Using machine learning methodologies, we evaluate the performance of the features derived from the two approaches. Our results show that head motion patterns are effective behavioural cues for detecting depression. Additionally, explanatory class-specific kinemes patterns can be observed, in alignment with prior research.

This paper makes the following research contributions:

- A study of head movements as a biomarker for clinical depression, which so far has been understudied.
- Proposing the *kineme* representation of motion patterns as an effective and explanatory means for depression analysis.
- A detailed investigation of various classifiers for 2-class and 4-class categorisation on the *AVEC2013* and *BlackDog* datasets. We obtain peak F1-scores of 0.79 and 0.82, respectively, on *thin-slice* chunks for binary classification on the *BlackDog* and *AVEC2013* datasets, which compare favorably to prior approaches. Also, a video-level F1-score of 0.72 is achieved for 4-class categorisation on *AVEC2013*.

The remainder of this paper is organised as follows. Section 2 provides an overview of related work. Section 3 describes the kineme formulation, followed by Section 4 that details the explainable kineme features used as a representation of motion patterns. The methodology is presented in Section 5, while Section 6 provides details of the datasets, experimental settings, and classifiers used in this study. The experimental results are shown and discussed in Section 7. Finally, the conclusions are drawn in Section 8.

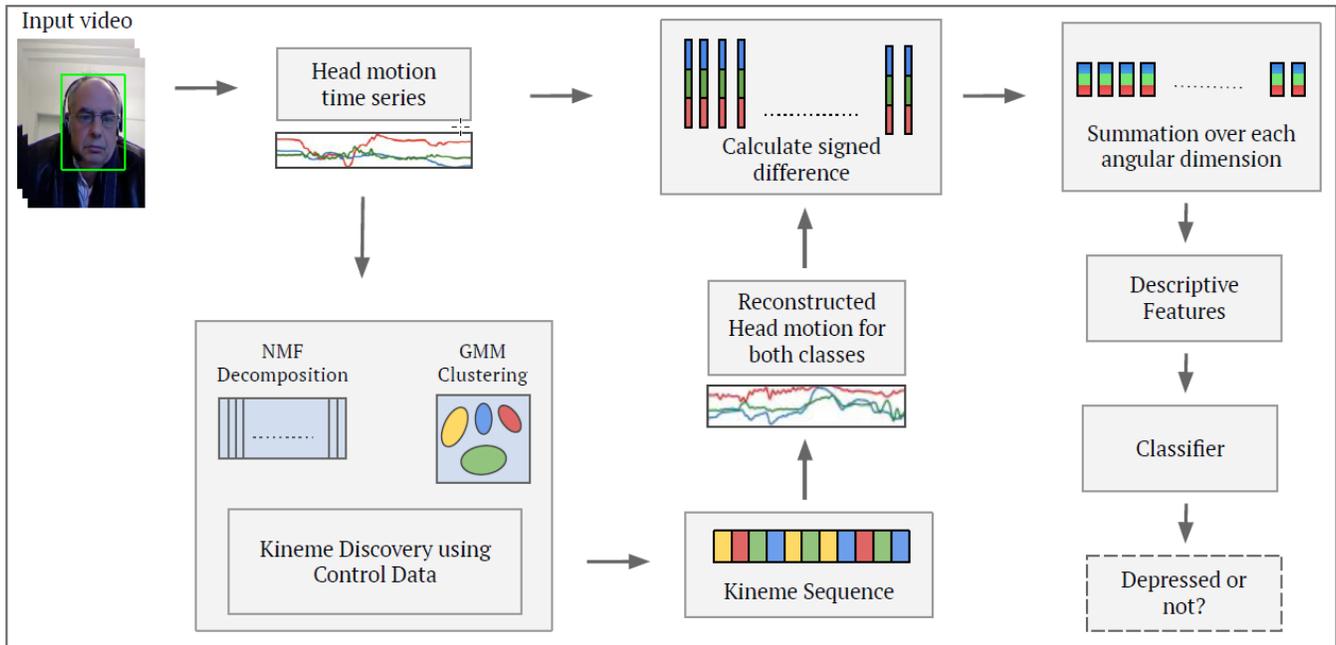


Figure 1: Overview: We learn kinemes for the control class, and the reconstruction errors between the raw and reconstructed head-motion segments, obtained via kineme clustering, are computed for both the control and depressed classes. Statistical descriptors over the yaw, pitch and roll dimensions (a total of 8×3 features) are utilized for depression detection via machine learning techniques.

2 RELATED WORK

In this section, we briefly review the literature focusing on (a) depression detection as a classification problem, and (b) depression detection using head motion patterns.

2.1 Depression Analysis as a Classification Task

Traditionally, depression detection has been approached as a supervised binary classification task, with many studies relying on discriminative classifiers to distinguish between *healthy controls* and *patients* [3, 5, 11]. A typical recognition accuracy of up to 80% demonstrates the promise of behavioural cues such as eye-blink and closed-eye duration rate, statistical features computed over the yaw, pitch and roll head-pose features, *etc.* to differentiate the two classes. However, challenges involved in depression detection such as limited clinically validated, curated data and a skewed data distributions have been acknowledged in the literature [5, 30].

Recent efforts have sought to learn patterns indicative of only the target class and reformulate depression detection as a one-class classification problem to mitigate the issues with imbalanced datasets [1, 32]. Studies have attempted to learn features associated with control participants and treat inputs that deviate from these patterns as *anomalous* [17, 29]. Gerych *et al.* [17] formulate the task as anomaly detection by leveraging autoencoders to learn features of the non-depressed class and treating depressed user data as outliers. Similarly, Mourão-Miranda *et al.* [29] employ a one-class SVM to classify patients as outliers compared to healthy participants based on the fMRI responses to sad facial expressions. Conversely, a few studies explore one-class classification by learning features characterising the depressed class, and treating non-depressed subjects as outliers [1, 32].

2.2 Depression Detection via Head Motion Cues

Many studies have focused on non-verbal behavioural cues, such as body gestures [23, 24], facial expressions [9, 13, 21], their combination [34] and speech features [12, 22, 36] as biomarkers for depression diagnosis and rehabilitation utilising computational tools [37]. Head motion patterns have nevertheless received little attention. Psychological research on depression assessment has identified head motion as a significant non-verbal cue for depression with more pronounced behavioural changes in hand and head regions as compared to other body parts for depressed patients [35]. Waxer *et al.* [45] found that depressed subjects are more likely to keep their heads in a downward position and exhibit significantly reduced head nodding compared to healthy subjects [16]. Another study focusing on social interactions identified the reduced involvement of depressed patients in conversations, where their behaviour was characterised by lesser encouragement (head nodding and backchanneling while listening) and fewer head movements [20].

From a computational standpoint, only a few studies have employed head pose and movement patterns for automatic depression detection. Alghowinem *et al.* [5] analysed head movements by modelling statistical features extracted from the 2D Active Appearance Model (AAM) projection of a 3D face and demonstrated the efficacy of head pose as behavioural cue. Another study [24] generated a histogram of head movements normalised over time to highlight the diminished movements of depressed patients due to psychomotor retardation, characterised by a more frequent occurrence of static head positions than in healthy controls. Several studies [11, 14, 28, 40] explored the utilisation of head motion as a complementary cue to other modalities to enhance detection performance. For instance, several studies [4, 6] combined head pose with

speech behaviour and eye gaze to develop statistical features for depression analysis. Generalisation across different cross-cultural datasets was attempted in [3] by using head pose and eye gaze based temporal features. Kacem *et. al.* [25] encoded head motion dynamics with facial expressions to classify depression based on severity, while Dibeklioglu *et. al.* [15] included vocal prosody in combination with head and facial movements for depression detection.

2.3 Novelty of the Proposed Approach

From the literature review, it can be seen that while a number of studies have employed head movements as a complementary cue in multimodal approaches, only few studies have deeply explored head motion as a rich source of information. Further, the explainability of behavioural features, especially head motion features, for depression detection has not yet been explored in the literature. This study (a) is the first to propose the use of kinemes as depression biomarkers, (b) explores multimodal cues derived from head motion behaviour as potential biomarkers for depression; specifically, we show that kinemes learned for the depressed and control classes, or only the control class enable accurate depression detection, and (c) the learned kinemes also *explain* depressed behaviours consistent with prior observations.

3 KINEME FORMULATION

This section describes our approach to discovering a set of elementary head motion units termed *kinemes* from 3D head pose angles. These head pose angles are expressed as a time-series of short overlapping segments, which enables shift invariance. The segments are then projected onto a lower-dimensional space and clustered using a Gaussian Mixture Model [38].

We extracted 3D head pose angles using the OpenFace tool [7] in terms of 3D Euler rotation angles, *pitch* (θ_p), *yaw* (θ_y) and *roll* (θ_r). The head movement over a duration T is denoted as a time-series: $\theta = \{\theta_p^{1:T}, \theta_y^{1:T}, \theta_r^{1:T}\}$. We ensure that the rotation angles remain non-negative by defining the range in $[0^\circ, 360^\circ]$.

For each video, the multivariate time-series θ is divided into short overlapping segments of length ℓ with overlap $\ell/2$, where the i^{th} segment is represented as a vector $\mathbf{h}^{(i)} = [\theta_p^{i:i+\ell}, \theta_y^{i:i+\ell}, \theta_r^{i:i+\ell}]$. Considering the total number of segments in any given video as s , the characterisation matrix \mathbf{H}_θ for this video is defined as $\mathbf{H}_\theta = [\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \dots, \mathbf{h}^{(s)}]$. Thus, for a training set of N samples, the head motion matrix is created as $\mathbf{H} = [\mathbf{H}_{\theta_1} | \mathbf{H}_{\theta_2} | \dots | \mathbf{H}_{\theta_N}]$ with each column of \mathbf{H} representing a single head motion time-series for a given video sample. We decompose $\mathbf{H} \in \mathbb{R}_+^{m \times n}$ into a basis matrix $\mathbf{B} \in \mathbb{R}_+^{m \times q}$ and a coefficient matrix $\mathbf{C} \in \mathbb{R}_+^{q \times n}$ using Non-negative Matrix Factorization (NMF) such that $m = 3\ell$, $n = Ns$

$$\min_{\mathbf{B} \geq 0, \mathbf{C} \geq 0} \|\mathbf{H} - \mathbf{BC}\|_F^2 \quad (1)$$

where $q \leq \min(m, n)$ and $\|\cdot\|_F$ denotes the Frobenius norm. Rather than clustering the raw head motion segments, we employ a more interpretable and stable approach by clustering the coefficient vectors in the transformed space. To this end, we learn a Gaussian Mixture Model (GMM) using the columns of the coefficient matrix \mathbf{C} to produce a $\mathbf{C}^* \in \mathbb{R}_+^{q \times k}$ where $k \ll Ns$. These vectors in the learned subspace are transformed back to the original head motion

subspace defined by the Euler angles using $\mathbf{H}^* = \mathbf{BC}^*$. The columns of matrix \mathbf{H}^* represent the set of K kinemes as $\{\mathcal{K}_i\}_{i=1}^K$.

Now, we can represent any head motion time-series θ as a sequence of kinemes discovered from the input video set by associating each segment of length ℓ from θ with one of the kinemes. For each i^{th} segment in the time-series, we compute the characterisation vector $\mathbf{h}^{(i)}$ and project it onto the transformed subspace defined by \mathbf{B} to yield $\mathbf{c}^{(i)}$ such that:

$$\hat{\mathbf{c}} = \arg \min_{\mathbf{c}^{(i)} \geq 0} \|\mathbf{h}^{(i)} - \mathbf{Bc}^{(i)}\|_F^2 \quad (2)$$

We then maximise the posterior probability $P(K|\hat{\mathbf{c}})$ over all kinemes to map the i^{th} segment with its corresponding kineme $K^{(i)}$. In the same way, we compute the corresponding kineme label for each segment of length ℓ to obtain a sequence of kinemes: $\{K^{(1)} \dots K^{(s)}\}$, where $K^{(j)} \in \mathcal{K}$ for all segments of time-series θ .

4 EXPLAINABLE KINEME FEATURES

We now examine kineme patterns obtained from the depression datasets, namely *BlackDog* [4] and *AVEC2013* [43] (described in Sec. 6.1). Using the *Openface* [7] toolkit, we extracted *yaw*, *pitch* and *roll* angles per frame, and segmented each video into 2s and 5s-long chunks with 50% overlap for the AVEC2013 and BlackDog datasets, respectively. Considering $K = 16$ [38], we extracted kinemes from both patient and healthy control segments, following the procedure outlined in Sec. 3. We further examined the kinemes learned for each dataset to identify the set of distinctive kinemes for the two classes. To obtain the most discriminative kinemes, we computed the relative frequency of occurrence for each kineme for the control and patient data, and selected the top five kinemes per class based on their relative frequency difference (see Sec. 5.1).

Selected kinemes corresponding to the maximal difference in their relative frequency of occurrence for the control and patient classes are visualised in Figures 2 (*BlackDog*) and 3 (*AVEC2013*). Examining the control-specific kinemes in Figs. 2 and 3, we observe a greater degree of movement for healthy subjects as compared to a predominantly static head pose conveyed by the depressed patient-specific kinemes. Head nodding, characterised by pitch oscillations, and considerable roll angle variations can be noted for at least one control-class kineme; conversely, patient-specific kinemes exhibit relatively small changes over all head pose angular dimensions. These findings are reflective of reduced head movements in the depressed cohort compared to healthy individuals, which is consistent with observations made in past studies [5, 20].

5 CLASSIFICATION METHODOLOGY

In this section, we outline the methodology for discovering kinemes from short overlapping video segments. Initially, we discovered kinemes utilising data segments from both control and patient classes (two-class kineme discovery or 2CKD approach; see Section 4). Subsequently, we learned kinemes solely from the healthy control cohort and utilised them to represent the head pose data of depressed patients (denoted as healthy control kineme discovery or HCKD) approach.

Given a time series θ , we divide it into short overlapping segments of uniform length, with a segment duration of 5s for the BlackDog and 2s for the AVEC dataset. These segment lengths

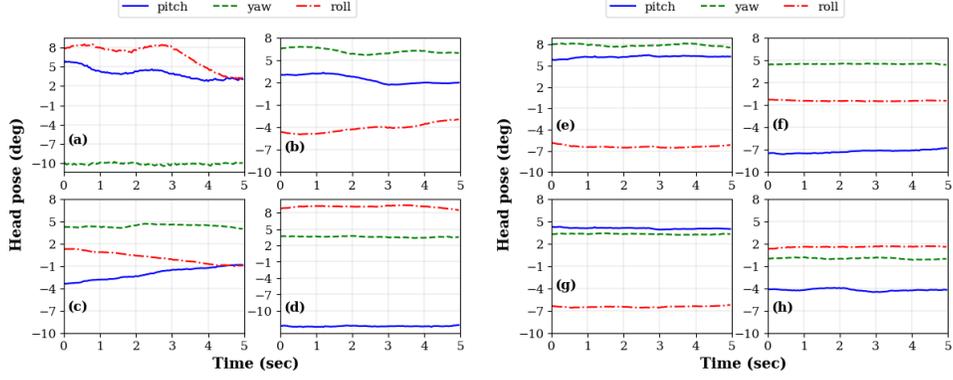


Figure 2: Plots of kinemes that occur more frequently for the control (left) and patient (right) cohorts in the *BlackDog* dataset.

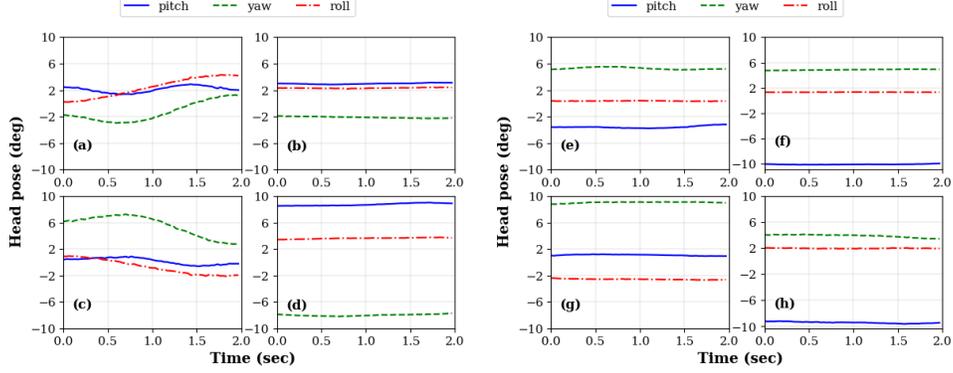


Figure 3: Plots of kinemes that occur more frequently for the minimally depressed (left) and patient (right) cohorts in *AVEC2013*.

were empirically chosen and provided the best results from among segment lengths spanning 2s to 7s for both datasets. For both approaches, a total of $K = 16$ kinemes are learned from the two datasets as per the procedure outlined in Section 3.

5.1 Kineme Discovery from Two-class Data

To examine whether the kinemes discovered from head pose angles of both classes are effective cues for depression detection, we learn kinemes from segments corresponding to both patient and control videos. Upon discovering the kineme values, the *relative frequency* η_{K_i} of each kineme K_i is computed over the two classes as:

$$\eta_{K_i} = \frac{f(K_i)}{\sum_{i=1}^{16} f(K_i)} \quad (3)$$

where $f(K_i)$ represents the frequency of occurrence of the kineme K_i for a particular class. We then compute the relative frequency difference for each kineme between the two classes to identify the ten most differentiating kinemes (four kinemes per class are depicted in Figs. 2, 3). Next, we generate a feature set by extracting the frequencies of the selected kinemes over the thin-slice chunks considered for analysis. Thus, we obtain a 10-dimensional feature vector representing kineme frequencies for each chunk.

5.2 Kineme Discovery from Control Data

Here, we learn kinemes representing head motion solely from the control cohort. Subsequently, head pose segments from both the patient and control classes are represented via the discovered kinemes, and reconstruction errors computed. Let the raw head pose vector

$\mathbf{h}^{(i)}$ for the i^{th} segment in the original subspace be denoted as:

$$\mathbf{h}^{(i)} = [\theta_p^{i:i+\ell} \ \theta_y^{i:i+\ell} \ \theta_r^{i:i+\ell}] \quad (4)$$

Let the kineme value associated with this segment be $K^{(i)}$. Based on the kinemes discovered from the control cohort alone, we calculate the reconstructed kineme for the i^{th} segment as $\tilde{\mathbf{h}}^{(i)}$. The reconstructed vector for each kineme is determined by converting the GMM cluster centre for each kineme from the learned space to the original *pitch-yaw-roll* space. The reconstructed head pose vector for the segment is:

$$\tilde{\mathbf{h}}^{(i)} = [\tilde{\theta}_p^{i:i+\ell} \ \tilde{\theta}_y^{i:i+\ell} \ \tilde{\theta}_r^{i:i+\ell}] \quad (5)$$

To compute the reconstruction error for both depressed patients and healthy controls, we compute the signed difference between the two vectors for each segment to account for the difference between raw head pose vector and the GMM cluster centres. We calculate the difference vector $\mathbf{d}^{(i)}$ for each i^{th} segment as:

$$\mathbf{d}^{(i)} = \mathbf{h}^{(i)} - \tilde{\mathbf{h}}^{(i)} = [d_p^{i:i+\ell} \ d_y^{i:i+\ell} \ d_r^{i:i+\ell}] \quad (6)$$

These signed differences values are added over each angular dimension of pitch (p), yaw (y), and roll (r) for the segment.

$$s_e^{(i)} = \sum_{n=1}^{\ell} d_e^{i:i+n} \quad (7)$$

where each $s_e^{(i)}$ is calculated for each angular dimension $e \in \{p, y, r\}$ over all segments of both classes. Depending on the thin-slice chunk duration considered for classification, we compute different descriptive statistics to generate the feature set. Considering number of elementary kineme chunks in the considered time-window to be n_c , we obtain the following feature vector for each angle $e \in \{p, y, r\}$:

$$\mathbf{as}_e = [|s_e^{(1)}|, |s_e^{(2)}|, \dots, |s_e^{(n_c)}|] \quad (8)$$

where $|\cdot|$ represents the absolute value. We then calculate eight statistical features from the above vectors, namely, *minimum*, *maximum*, *range*, *mean*, *median*, *standard deviation*, *skewness*, and *kurtosis* (total of 8×3 features over the yaw, pitch, roll dimensions).

6 EXPERIMENTS

We perform binary classification on the BlackDog and AVEC2013 datasets, plus 4-class classification on AVEC2013. This section details our datasets, experimental settings and learning algorithms.

6.1 Datasets

We examine two datasets in this study: clinically validated data collected at the Black Dog Institute – a clinical research facility focusing on the diagnosis and treatment of mood disorders such as anxiety and depression (referred to as *BlackDog* dataset) – and the *Audio/Visual Emotion Challenge* (AVEC2013) depression dataset.

BlackDog Dataset [4]: This dataset comprises responses from healthy controls and depression patients selected as per the criteria outlined in the Diagnostic and Statistic Manual of Mental Disorders (DSM-IV). Healthy controls with no history of mental illness and patients diagnosed with severe depression were carefully selected [4]. For our analysis, we focus on the structured interview responses in [4], where participants answered open-ended questions about life events, designed to elicit spontaneous self-directed responses, asked by a clinician. In this study, we analyse video data from 60 subjects (30 depressed patients and 30 healthy controls), with interview durations ranging from 183 – 1200s.

AVEC2013 Dataset [43]: Introduced for a challenge in 2013, this dataset is a subset of the audio-visual depressive language corpus (AViD-corpus) comprising 340 video recordings of participants performing different PowerPoint guided tasks detailed in [43]. The videos are divided into three nearly equal partitions (training, development, and test) with videos ranging from 20 – 50min. Each video frame depicts only one subject, although some participants feature in multiple video clips. The participants completed a multiple-choice inventory based on the Beck Depression Index (BDI) [8] with scores ranging from 0 to 63 denoting the severity of depression. For binary classification, we dichotomise the recordings into the non-depressed and depressed cohorts as per the BDI scores. Subjects with a BDI score ≤ 13 are categorised as *non-depressed*, while the others are considered as *depressed*.

AVEC2013 Multi-Class Classification: For fine-grained depression detection over the AVEC2013 dataset, we categorise the dataset based on the BDI score into four classes as detailed below:

- Nil or minimal depression: BDI score 0 - 13
- Mild depression: BDI score 14 - 19
- Moderate depression: BDI score 20 - 28
- Severe depression: BDI score 29 - 63

6.2 Experimental Settings

Implementation Details: For binary classification, we evaluate performance for the smaller *BlackDog* dataset via 5-repetitions of 10-fold cross-validation (10FCV). For the AVEC2013, the pre-partitioned train, validation and test sets are employed. We utilise the validation sets for fine-tuning classifier hyperparameters.

Chunk vs Video-level Classification: The videos from both datasets are segmented into smaller chunks of 15s – 135s length, to examine the influence of *thin-slice* chunk duration on the classifier performance. We repeated the video label for all chunks and metrics are computed over all chunks for chunk-level analysis. Additionally, video-level classification results are obtained by computing the majority label over all video chunks in the test set.

Performance Measures: For the BlackDog dataset, results are shown as $\mu \pm \sigma$ values over 50 runs (5 \times 10FCV repetitions). For AVEC2013, performance on the test set is reported. For both, we evaluate performance via the accuracy (Acc), weighted F1 (F1), precision (Pr), and recall (Re) metrics. The weighted F1-score denotes the mean F1-score over the two classes, weighted by class size.

6.3 Classification Methods

Given that our proposed features do not model spatial or temporal correlations, we employ different machine learning models for detecting depression as described below:

- **Logistic Regression (LR)**, a probabilistic classifier that employs a sigmoid function to map input observations to binary labels. We utilise extensive grid-search to fine-tune parameters such as penalty $\in \{l1, l2, None\}$ and regulariser $\lambda \in \{1e^{-6}, \dots, 1e^3\}$.
- **Random Forest (RF)**, where multiple decision trees are generated from training data whose predictions are aggregated for labelling. Fine-tuned parameters include the number of estimators $N \in [2, \dots, 8]$, maximum depth $\in [3, \dots, 7]$, and maximum features in split $\in [3, \dots, 7]$.
- **Support Vector Classifier (SVC)**, a discriminative classifier that works by transforming training data to a high-dimensional space where the two classes can be linearly separated via a hyperplane. For SVC, we examine different kernels $\in \{rbf, poly, sigmoid\}$ and fine-tune regularisation parameter $C \in \{0.1, 1, 10, 100\}$ and kernel coefficient $\gamma \in \{0.0001, \dots, 1, scale, auto\}$.
- **Extreme Gradient Boosting (XGB)**, a model built upon a gradient boosting framework, and focused on improving a series of weak learners by employing the gradient descent algorithm in a sequential manner. The fine-tuned hyperparameters include the number of estimators $\{50, 100, 150\}$, maximum depth $\in [3, \dots, 7]$ of the tree and learning rate $\in [0.0005, \dots, 0.1]$.
- **Multi Layer Perceptron (MLP)**, where we employed a feed-forward neural network with two hidden dense layers comprising 12 and 6 neurons, resp., with a rectified linear unit (ReLU) activation. For training, we employ categorical cross-entropy as the loss function and fine-tune the following hyperparameters: learning rate $\in \{1e^{-4}, 1e^{-3}, 1e^{-2}\}$, and batch size $\in \{16, 24, 32, 64\}$. We utilise the Adam optimiser for updating the network weights during training.

7 RESULTS AND DISCUSSION

Table 1 shows the classification results obtained for the *BlackDog* dataset with the 2CKD and HCKD approaches (Section 5). Table 2 presents the corresponding results for the *AVEC2013* dataset. These tables present classification measures obtained at the *chunk-level* (best results achieved over 15–135s-long chunks for the two datasets are presented), and the *video-level* (label derived upon computing the mode over the chunk-level labels). Based on these results, we make the following observations:

- It can be noted from Tables 1 and 2 that relatively lower accuracies and F1 scores are achieved for both datasets using the 2CKD approach, implying that while class-characteristic kinemes are explanative as seen from Figs. 2 and 3, they are nevertheless not discriminative enough to effectively distinguish between the two classes.
- In comparison, we note far superior performance with the HCKD method over all classifiers. As a case in point, we obtain peak chunk-level F1-scores of 0.79 and 0.62, resp., for HCKD and 2CKD on BlackDog, while the corresponding F1-scores are 0.82 and 0.61, resp., on AVEC. This observation reveals considerable and distinguishable differences in the reconstruction errors for the patient and control classes, and convey that patient data are characterised as *anomalies* when kinemes are only learned from the control cohort.
- Examining the HCKD precision and recall measures for both datasets, we note higher precision than recall at the chunk-level for the BlackDog dataset. Nevertheless, higher recall is achieved at the video-level with multiple classifiers. Likewise, higher chunk-level precision is noted for AVEC, even if ceiling video-level precision and recall are achieved.
- Comparing HCKD chunk and video-level F1-scores for both datasets, similar or higher video-level F1 values can be seen in Table 1. F1-score differences are starker in Table 2, where video-level scores are considerably higher than chunk-level scores. These results suggest that aggregating observations over multiple thin-slice chunks is beneficial and enables more precise predictions as shown in [27].
- Examining measures achieved with the different classifiers, the support vector classifier achieves the best chunk-level F1-score on both datasets, with the LR classifier performing very comparably. All classifiers achieve very similar performance when video-level labels are compared.

7.1 Comparison with the state-of-the-art

Our best results are compared against prior classification-based depression detection studies in Table 3. For the *BlackDog* dataset, Alghowinem *et al.* [5] analysed statistical functional features extracted from a 2D Active Appearance Model, whereas Joshi *et al.* [24] computed a histogram of head movements by estimating the displacement of fiducial facial points. Compared to N -average recall of 0.71 reported in [5], and an accuracy of 0.72 noted in [24], our kineme-based approach achieves better chunk and video-level accuracies (0.75 and 0.80, resp.), and superior chunk-level recall (0.81). As most previous studies on the *AVEC2013* dataset focus on continuous prediction, we compare our model’s performance with the *AVEC2014* [42] results examining visual features.

AVEC2014 used the same subjects as *AVEC2013*, but with additional, specific task data (*Northwind*, *Freeform*) extracted from the *AViD* videos. For video analysis, Senoussaoui *et al.* [39] extracted LGBP-TOP features from frame blocks to obtain an accuracy of 0.82 using an SVM classifier. On the other hand, Al-gawwam *et al.* [2] extracted eye-blink features from video data using a facial landmark tracker to achieve an accuracy of 0.92 for the *Northwind* task and 0.88 for the *Freeform* task. Comparatively, our work achieves an accuracy of 0.82 at the chunk-level and 1.00 at the video-level. The next section will detail the performance of a more fine-grained 4-class categorisation on the *AVEC2013* dataset.

7.2 AVEC2013 Multi-class Classification

Table 4 depicts video-level 4-class classification results achieved on the *AVEC2013* dataset via the HCKD approach. The 4-class categorisation was performed to further validate the correctness of the HCKD approach, which produces ceiling video-level F1, Precision and Recall measures on *AVEC2013* in binary classification. Results are reported on the test set, upon fine-tuning the classifier models on the development set. Reasonably good F1-scores are achieved even with 4-class classification, with a peak F1 of 0.72 obtained with the LR, RF and support vector classifiers. Cumulatively, our empirical results confirm that kinemes encoding atomic head movements are able to effectively differentiate between (a) the patient and control classes, and (b) different depression severity bands.

7.3 Ablative Analysis over Thin Slices

Tables 1 and 2 evaluate detection performance over (*thin-slice*) chunks or short behavioural episodes, and over the entire video, on the *BlackDog* and *AVEC2013* datasets. We further compared labelling performance at the chunk and video-levels using chunks spanning 15 – 135s. The corresponding results are presented in Figure 4. For both plots presented in the figure, the dotted curves denote video-level F1-scores, while solid curves denote chunk-level scores obtained for different classifiers.

For the *BlackDog* dataset (Fig. 4 (left)), longer time-slices (of length 75 – 105s) achieve better performance than shorter (15 – 60s long) ones at both the chunk and video-levels across all classifiers; these findings are consistent with the finding that more reliable predictions can be achieved with longer observations in general [27]. However, a performance drop is noted for very long chunk-lengths of 120 – 135s duration. Decoding results on the *AVEC2013* dataset, consistent with Table 3 results, a clear gap is noted between the chunk and video-level results, with the latter demonstrating superior performance. Very similar F1-scores are observed across classifiers for various chunk lengths. No clear trends are discernible from video-level F1-scores obtained with different chunk-lengths, except that the performance in general decreases for all classifiers with very long chunks.

7.4 Ablative Analysis over Angular Dimensions

To investigate the impact of the head pose angular dimensions on chunk-level binary depression detection performance, we perform detection utilising (8×1) statistical features over each of the (pitch, yaw, and roll) angular dimensions, and concatenate features (8×2)

Table 1: Chunk and Video-level classification results on the BlackDog dataset with the 2CKD and HCKD approaches. Accuracy (Acc), F1, Precision (Pr) and Recall (Re) are tabulated as ($\mu \pm \sigma$) values.

Condition	Classifier	Chunk-level				Video-level			
		Acc	F1	Pr	Re	Acc	F1	Pr	Re
2CKD	LR	0.60±0.15	0.61±0.14	0.67±0.22	0.65±0.22	0.60±0.20	0.59±0.21	0.55±0.30	0.65±0.33
	RF	0.58±0.13	0.60±0.12	0.67±0.21	0.61±0.21	0.61±0.19	0.62±0.19	0.59±0.32	0.59±0.32
	SVC	0.60±0.15	0.62±0.15	0.68±0.25	0.62±0.25	0.62±0.19	0.63±0.19	0.61±0.32	0.59±0.33
	XGB	0.55±0.17	0.54±0.16	0.63±0.21	0.71±0.22	0.53±0.17	0.50±0.20	0.54±0.23	0.79±0.21
	MLP	0.53±0.15	0.52±0.17	0.60±0.22	0.71±0.21	0.51±0.20	0.47±0.21	0.53±0.27	0.74±0.32
HCKD	LR	0.77±0.13	0.78±0.12	0.85±0.19	0.74±0.21	0.79±0.16	0.78±0.17	0.81±0.30	0.66±0.31
	RF	0.71±0.13	0.73±0.12	0.75±0.25	0.71±0.20	0.76±0.15	0.76±0.16	0.75±0.26	0.82±0.25
	SVC	0.78±0.14	0.79±0.13	0.87±0.18	0.74±0.20	0.80±0.18	0.80±0.19	0.83±0.30	0.70±0.31
	XGB	0.72±0.13	0.72±0.12	0.75±0.18	0.81±0.15	0.78±0.17	0.78±0.17	0.74±0.27	0.82±0.27
	MLP	0.75±0.13	0.76±0.12	0.78±0.21	0.81±0.21	0.76±0.16	0.76±0.16	0.74±0.29	0.77±0.28

Table 2: Chunk and Video-level classification results on the AVEC2013 dataset with the 2CKD and HCKD approaches. Accuracy (Acc), F1, Precision (Pr) and Recall (Re) are tabulated as ($\mu \pm \sigma$) values.

Condition	Classifier	Chunk-level				Video-level			
		Acc	F1	Pr	Re	Acc	F1	Pr	Re
2CKD	LR	0.58	0.58	0.54	0.65	0.61	0.61	0.57	0.71
	RF	0.61	0.61	0.57	0.59	0.72	0.72	0.67	0.82
	SVC	0.61	0.61	0.57	0.63	0.64	0.64	0.61	0.65
	XGB	0.59	0.58	0.57	0.44	0.67	0.67	0.65	0.65
	MLP	0.56	0.56	0.52	0.60	0.58	0.58	0.65	0.65
HCKD	LR	0.80	0.80	0.77	0.81	0.94	0.94	0.94	0.94
	RF	0.78	0.78	0.78	0.75	1.00	1.00	1.00	1.00
	SVC	0.82	0.82	0.83	0.77	1.00	1.00	1.00	1.00
	XGB	0.80	0.80	0.79	0.77	1.00	1.00	1.00	1.00
	MLP	0.81	0.80	0.80	0.77	1.00	1.00	1.00	1.00

Table 3: Comparison with prior works for the two datasets.

Dataset	Methods	Features	Evaluation metrics			
			Acc	F1	Pr	Re
BlackDog	Alghowinem <i>et al.</i> [5]	Head movement	-	-	-	0.71
	Joshi <i>et al.</i> [24]	Head movement	0.72	-	-	-
	Ours (Chunk-level)	Kinemes	0.75	0.76	0.78	0.81
	Ours (Video-level)	Kinemes	0.80	0.80	0.83	0.70
AVEC2013	Senoussaoui <i>et al.</i> (AVEC2014) [39]	Video features	0.82	-	-	-
	Al-gawwam <i>et al.</i> (AVEC2014 - Northwind) [2]	Eye Blink	0.85	-	-	-
	Al-gawwam <i>et al.</i> (AVEC2014 - Freeform) [2]	Eye Blink	0.92	-	-	-
	Ours (AVEC2013 at chunk-level)	Kinemes	0.82	0.82	0.83	0.87
	Ours (AVEC2013 at Video-level)	Kinemes	1.00	1.00	1.00	1.00

Table 4: Video-level 4-class categorization results on the AVEC dataset obtained with the HCKD approach. Accuracy (Acc), F1, Precision (Pre) and Recall (Re) are tabulated.

Condition	Classifier	Video-level			
		Acc	F1	Pr	Re
HCKD	LR	0.71	0.72	0.73	0.71
	RF	0.74	0.72	0.80	0.74
	SVC	0.74	0.72	0.75	0.74
	XGB	0.71	0.69	0.68	0.71
	MLP	0.69	0.66	0.64	0.69

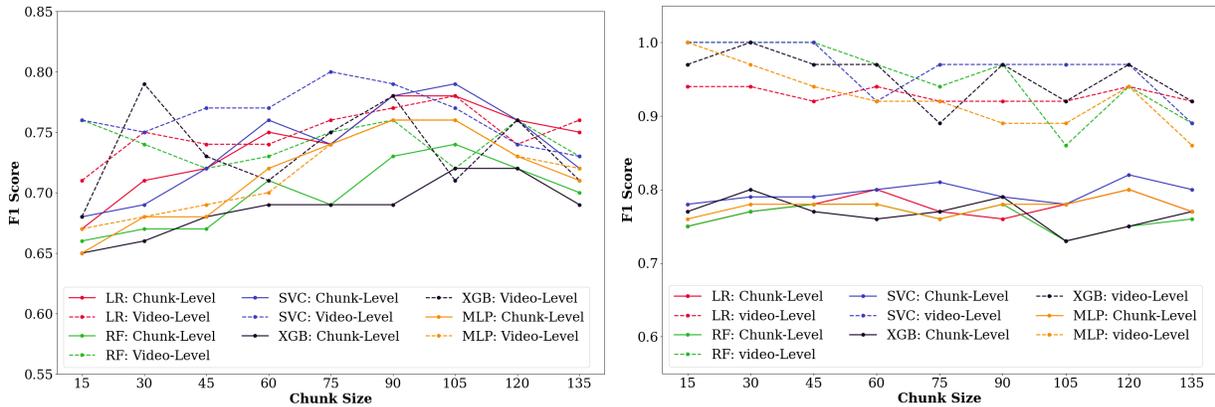


Figure 4: Chunk vs video-level performance comparison for the BlackDog (left) and AVEC2013 (right) datasets.

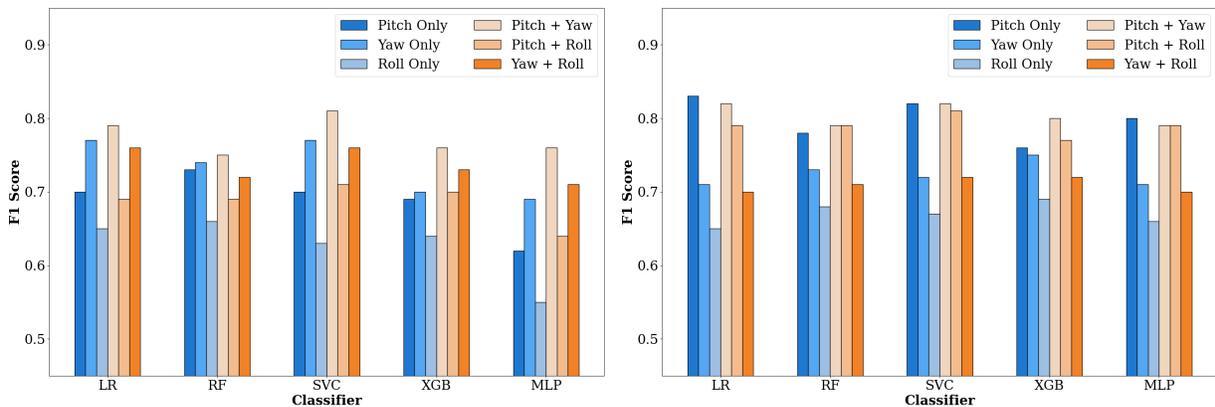


Figure 5: Comparing F1 scores with different descriptors for the BlackDog (left) and AVEC2013 (right) datasets across classifiers.

for the dimensional pairs to evaluate which angular dimension(s) are more informative.

Figure 5 presents F1-scores obtained with the different classifiers for uni-dimensional and pairwise-dimensional features. On the BlackDog dataset, a combination of the pitch and yaw-based descriptors produce the best performance across all models, while roll-specific descriptors perform worst. For the AVEC2013 dataset, pitch-based descriptors achieve excellent performance across models. The F1-scores achieved with these features are very comparable to the pitch + yaw and pitch + roll combinations. Here again, roll-specific features achieve the worst performance. Cumulatively, these results convey that pitch is the most informative head pose dimension, with roll being the least informative. With respect to combinations, the pitch + yaw combination in general produces the best results. These results again confirm that responsiveness in social interactions, as captured by pitch (capturing actions such as head nodding) and yaw (capturing head shaking), provides a critical cue for detecting depression, consistent with prior studies [5, 20].

8 CONCLUSION

In this paper, we demonstrate the efficacy of elementary head motion units, termed *kinemes*, for depression detection by utilising two approaches: (a) discovering kinemes from data of both patient and control cohorts, and (b) learning kineme patterns solely from

the control cohort to compute statistical functional features derived from reconstruction errors for the two classes. Apart from effective depression detection, we also identify explainable kineme patterns for the two classes, consistent with prior research.

Our study demonstrates the utility of head motion features for detecting depression, but our experiments are restricted to classification tasks involving a discretisation of the depression scores. In the future, we will investigate (a) the utility of kinemes for continuous prediction (regression) of depression severity, (b) the cross-dataset generalisability of models trained via kinemes, and (c) the development of multimodal methodologies combining kinemes with other behavioural markers, and evaluating their efficacy.

REFERENCES

- [1] Juan Aguilera, Delia Irazú Hernández Farías, Rosa María Ortega-Mendoza, and Manuel Montes-y Gómez. 2021. Depression and anorexia detection in social media as a one-class classification problem. *Applied Intelligence* 51 (2021), 6088–6103.
- [2] Sarmad Al-gawwam and Mohammed Benaissa. 2018. Depression detection from eye blink features. In *2018 IEEE international symposium on signal processing and information technology (ISSPIT)*. IEEE, 388–392.
- [3] Sharifa Alghowinem, Roland Goecke, Jeffrey F Cohn, Michael Wagner, Gordon Parker, and Michael Breakspear. 2015. Cross-cultural detection of depression from nonverbal behaviour. In *2015 11th IEEE International conference and workshops on automatic face and gesture recognition (FG)*, Vol. 1. IEEE, 1–8.
- [4] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear. 2016. Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors.

- IEEE Transactions on Affective Computing* 9, 4 (2016), 478–490.
- [5] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parker, and Michael Breakspear. 2013. Head pose and movement analysis as an indicator of depression. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 283–288.
 - [6] Sharifa Mohammed Alghowinem, Tom Gedeon, Roland Goecke, Jeffrey Cohn, and Gordon Parker. 2020. Interpretation of depression detection models via feature selection methods. *IEEE transactions on affective computing* (2020).
 - [7] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1–10. <https://doi.org/10.1109/WACV.2016.7477553>
 - [8] Aaron T Beck, Robert A Steer, Roberta Ball, and William F Ranieri. 1996. Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. *Journal of personality assessment* 67, 3 (1996), 588–597.
 - [9] Cecilia Bourke, Katie Douglas, and Richard Porter. 2010. Processing of facial emotion expression in major depression: a review. *Australian & New Zealand Journal of Psychiatry* 44, 8 (2010), 681–696.
 - [10] Antonio Campayo, Carlos H Gómez-Biel, and Antonio Lobo. 2011. Diabetes and depression. *Current psychiatry reports* 13, 1 (2011), 26–30.
 - [11] Jeffrey F Cohn, Nicholas Cummins, Julien Epps, Roland Goecke, Jyoti Joshi, and Stefan Scherer. 2018. Multimodal assessment of depression from behavioral signals. *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2* (2018), 375–417.
 - [12] Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke. 2011. An investigation of depressed speech detection: Features and normalization. In *Twelfth Annual Conference of the International Speech Communication Association*.
 - [13] Wheidima Carneiro de Melo, Eric Granger, and Abdenour Hadid. 2019. Combining global and local convolutional 3d networks for detecting depression from facial expressions. In *2019 14th IEEE international conference on automatic face & gesture recognition (fg 2019)*. IEEE, 1–8.
 - [14] Hamdi Dibeklioglu, Zakia Hammal, and Jeffrey F Cohn. 2017. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE journal of biomedical and health informatics* 22, 2 (2017), 525–536.
 - [15] Hamdi Dibeklioglu, Zakia Hammal, Ying Yang, and Jeffrey F Cohn. 2015. Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 307–310.
 - [16] Luciano Fossi, C Faravelli, and M Paoli. 1984. The ethological approach to the assessment of depressive disorders. *The Journal of nervous and mental disease* 172, 6 (1984), 332–341.
 - [17] Walter Gerych, Emmanuel Agu, and Elke Rundensteiner. 2019. Classifying depression in imbalanced datasets using an autoencoder-based anomaly detection approach. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*. IEEE, 124–127.
 - [18] Robert D Goldney, David Wilson, Eleonora Dal Grande, Laura J Fisher, and Alexander C McFarlane. 2000. Suicidal ideation in a random community sample: attributable risk due to depression and psychosocial and traumatic events. *Australian & New Zealand Journal of Psychiatry* 34, 1 (2000), 98–106.
 - [19] Paul E Greenberg, Andree-Anne Fournier, Tammy Sisitsky, Crystal T Pike, and Ronald C Kessler. 2015. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *The Journal of clinical psychiatry* 76, 2 (2015), 5356.
 - [20] William W Hale III, Jaap HC Jansen, Antoinette L Bouhuys, Jack A Jenner, and Rutger H van den Hoofdakker. 1997. Non-verbal behavioral interactions of depressed patients with partners and strangers: The role of behavioral social support and involvement in depression persistence. *Journal of affective disorders* 44, 2-3 (1997), 111–122.
 - [21] Lang He, Chenguang Guo, Prayag Tiwari, Hari Mohan Pandey, and Wei Dang. 2022. Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence. *International Journal of Intelligent Systems* 37, 12 (2022), 10140–10156.
 - [22] Zhaocheng Huang, Julien Epps, and Dale Joachim. 2019. Investigation of speech landmark patterns for depression detection. *IEEE Transactions on Affective Computing* 13, 2 (2019), 666–679.
 - [23] Jyoti Joshi, Abhinav Dhall, Roland Goecke, and Jeffrey F Cohn. 2013. Relative body parts movement for automatic depression analysis. In *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 492–497.
 - [24] Jyoti Joshi, Roland Goecke, Gordon Parker, and Michael Breakspear. 2013. Can body expressions contribute to automatic depression analysis?. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–7.
 - [25] Anis Kacem, Zakia Hammal, Mohamed Daoudi, and Jeffrey Cohn. 2018. Detecting depression severity by interpretable representations of motion dynamics. In *2018 13th IEEE international conference on automatic face & gesture recognition (fg 2018)*. IEEE, 739–745.
 - [26] Jean-Pierre Lépine and Mike Briley. 2011. The increasing burden of depression. *Neuropsychiatric disease and treatment* 7, sup1 (2011), 3–7.
 - [27] Surbhi Madan, Monika Gahalawat, Tanaya Guha, and Ramanathan Subramanian. 2021. Head Matters: Explainable Human-Centered Trait Prediction from Head Motion Dynamics. In *Proceedings of the 2021 International Conference on Multimodal Interaction (Montréal, QC, Canada) (ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 435–443. <https://doi.org/10.1145/3462244.3479901>
 - [28] Michelle Morales, Stefan Scherer, and Rivka Levitan. 2017. A cross-modal review of indicators for depression detection systems. In *Proceedings of the fourth workshop on computational linguistics and clinical psychology—From linguistic signal to clinical reality*. 1–12.
 - [29] Janaina Mourão-Miranda, David R Hardoon, Tim Hahn, Andre F Marquand, Steve CR Williams, John Shawe-Taylor, and Michael Brammer. 2011. Patient classification as an outlier detection problem: an application of the one-class support vector machine. *Neuroimage* 58, 3 (2011), 793–804.
 - [30] Md Nasir, Arindam Jati, Prashanth Gurnath Shivakumar, Sandeep Nalan Chakravarthula, and Panayiotis Georgiou. 2016. Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*. 43–50.
 - [31] Institute of Health Metrics and Evaluation. 2021. Global Health Data Exchange (GHDx).
 - [32] Kennedy Opoku Asare, Aku Visuri, and Denzil ST Ferreira. 2019. Towards early detection of depression through smartphone sensing. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 1158–1161.
 - [33] Anastasia Pampouchidou, Panagiotis G. Simos, Kostas Marias, Fabrice Meriaudeau, Fan Yang, Matthew Padiaditis, and Manolis Tsiknakis. 2019. Automatic Assessment of Depression Based on Visual Cues: A Systematic Review. *IEEE Transactions on Affective Computing* 10, 4 (2019), 445–470. <https://doi.org/10.1109/TAFFC.2017.2724035>
 - [34] Viral Parekh, Pin Sym Foong, Shengdong Zhao, and Ramanathan Subramanian. 2018. AVEID: Automatic Video System for Measuring Engagement In Dementia. In *23rd International Conference on Intelligent User Interfaces (Tokyo, Japan) (IUI '18)*. 409–413. <https://doi.org/10.1145/3172944.3173010>
 - [35] Jesper Pedersen, JTM Schelde, E Hannibal, K Behnke, BM Nielsen, and M Hertz. 1988. An ethological description of depression. *Acta psychiatrica scandinavica* 78, 3 (1988), 320–330.
 - [36] Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, Said Agrebi, and Alice Othmani. 2022. MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control* 71 (2022), 103107.
 - [37] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*. 3–12.
 - [38] Atanu Samanta and Tanaya Guha. 2017. On the role of head motion in affective expression. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2886–2890.
 - [39] Mohammed Senoussaoui, Milton Sarria-Paja, João F Santos, and Tiago H Falk. 2014. Model fusion for multimodal depression classification and level detection. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. 57–63.
 - [40] Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar. 2020. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing* 13, 2 (2020), 829–844.
 - [41] Annika Steffen, Julia Nübel, Frank Jacob, Jörg Bätzing, and Jakob Holstiege. 2020. Mental and somatic comorbidity of depression: a comprehensive cross-sectional analysis of 202 diagnosis groups using German nationwide ambulatory claims data. *BMC Psychiatry* (March 2020).
 - [42] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. AVEC 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*. 3–10.
 - [43] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. 3–10.
 - [44] Theo Vos, Christine Allen, and Megha Arora. 2016. 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* 388, 10053 (2016), 1545–1602. [https://doi.org/10.1016/S0140-6736\(16\)31678-6](https://doi.org/10.1016/S0140-6736(16)31678-6)
 - [45] Peter Waxer. 1974. Nonverbal cues for depression. *Journal of Abnormal Psychology* 83, 3 (1974), 319.