Automatic Assessment of Depression Based on Visual Cues: A Systematic Review

Anastasia Pampouchidou, Panagiotis G. Simos, Kostas Marias, Fabrice Meriaudeau, Fan Yang, Matthew Pediaditis, and Manolis Tsiknakis

Abstract—Automatic depression assessment based on visual cues is a rapidly growing research domain. The present exhaustive review of existing approaches as reported in over sixty publications during the last ten years focuses on image processing and machine learning algorithms. Visual manifestations of depression, various procedures used for data collection, and existing datasets are summarized. The review outlines methods and algorithms for visual feature extraction, dimensionality reduction, decision methods for classification and regression approaches, as well as different fusion strategies. A quantitative meta-analysis of reported results, relying on performance metrics robust to chance, is included, identifying general trends and key unresolved issues to be considered in future studies of automatic depression assessment utilizing visual cues alone or in combination with vocal or verbal cues.

Index Terms—Depression assessment, affective computing, facial expression, machine learning, facial image analysis

1 Introduction

The present work is a systematic review of existing methods for automatic detection and/or severity assessment of depression. Emphasis is given to approaches utilizing visual signs from the image processing and machine learning perspective in an attempt to fill the gap of previous comprehensive reviews. The aim of the review is to examine methods for automated depression analysis, which could assist clinicians in the diagnosis and monitoring of depression. The main questions addressed are whether: (a) video-based depression assessment can assist the diagnosis and monitoring of the disorder, and (b) if visual cues alone are sufficient or if they need to be supplemented by information from other modalities. State of the art methods are presented highlighting their advantages and limitations, based on a quantitative meta-analysis of their results. Datasets created to serve the various studies and the corresponding data acquisition protocols are also described and discussed.

1.1 Clinical Background of Depression

Depression is the most common mood disorder characterized by persistent negative affect [1]. Clinically distinct depressive disorders encompass a wide range of manifestations. According to the Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association (APA) [2], now in its fifth edition (DSM-5), subtypes of depressive disorders include: Major Depressive Disorder (MDD), Persistent Depressive Disorder (Dysthymia), Disruptive Mood Dysregulation Disorder (DMDD), Premenstrual Dysphoric Disorder (PDD), Substance/Medication-Induced Depressive Disorder (S/M-IDD), Depressive Disorder Due to Another Medical Condition (DDDAMC), and Other Specified Depressive Disorder (OSDD) or Unspecified Depressive Disorder (UDD).

According to DSM-5 MDD, commonly referred to as Clinical Depression, can be diagnosed by the presence of a) depressed mood most of the day, and/or b) markedly diminished interest or pleasure, combined with at least four of the following symptoms for a period exceeding two weeks:

- Significant weight change of over 5 percent in a month
- Sleeping disturbances (insomnia or hypersomnia)
- Psychomotor agitation or retardation almost every day
- Fatigue or loss of energy almost every day
- Feelings of worthlessness or excessive guilt
- Diminished ability to concentrate or indecisiveness almost every day
- Recurrent thoughts of death or suicidal ideation

- A. Pampouchidou and F. Yang are with the Le2i Laboratory, University of Burgundy, Le Creusot, Dijon 21078, France. E-mail: anastasia.pampouchidou@gmail.com, fanyang@u-bourgogne.fr.
- P.G. Simos is with the Division of Psychiatry, School of Medicine, University of Crete, Heraklion, Crete GR-70013, Greece. E-mail: akis.simos@gmail.com.
- K. Marias and M. Tsiknakis are with the Technological Educational Institute of Crete, Department of Informatics Engineering, Heraklion, Crete 714 10, Greece and with the Institute of Computer Science, Foundation for Research & Technology-Hellas, Heraklion, Crete GR-70013, Greece. E-mail: kmarias@ics.forth.gr, tsiknaki@ie.teicrete.gr.
- F. Meriaudeau is with the Le2i Laboratory, University of Burgundy, Le Creusot, Dijon 21078, France and with CISIR, Electrical Engineering DepartmentUniversiti Teknologi Petronas, Bota, Perak 32600, Malaysia. E-mail: fabrice.meriaudeau@utp.edu.my.
- M. Pediaditis is with the Institute of Computer Science, Foundation for Research & Technology-Hellas, Heraklion, Crete GR-70013, Greece. E-mail: matthew.pediaditis@gmail.com.

Manuscript received 7 Oct. 2015; revised 2 May 2017; accepted 28 June 2017. Date of publication 28 Sept. 2017; date of current version 25 Nov. 2019. (Corresponding author: Anastasia Pampouchidou.) Recommended for acceptance by M. Valstar. Digital Object Identifier no. 10.1109/TAFFC.2017.2724035

An additional common feature of all depressive disorders is "(...) *clinically significant distress or impairment in social, occupational, or other important areas of functioning* (...)" [2]. With MDD considered as the most typical form of the disease, other depressive disorders share some of the MDD symptoms, while each is distinguished by additional characteristics. For instance, chronicity of symptoms characterizes Dysthymia, also known as chronic depression. DMDD, also chronic, is characterized by severe persistent irritability and recurrent outbursts. PDD requires occurrence of depressive symptoms over a minimum of two menstrual cycles. The onset of depressive symptomatology should be clearly linked to persistent use or withdrawal from substances in order to justify diagnosis of S/M-IDD. Similarly, there should be a clear link between another serious medical condition and emergence of depressive symptomatology in DDDAMC. A diagnosis of Other Specified or Unspecified Depressive Disorders is reserved for cases where full criteria are not met for any of the aforementioned depressive disorders.

MDD is reported to be the fourth most prominent cause of disability and is expected to become the second by 2020 due to its increasing prevalence [1], [3], [4]. The "Survey of Health, Ageing and Retirement in Europe" [5] documents a consistent rise of depression among adults with increasing age, which is associated with significantly elevated risk for suicidal behavior [6]. The ongoing economic crisis in Europe resulting in high unemployment is implicated as a trigger, since 70-76 percent of unemployed people have been reported to display significant depressive symptomatology [7]. Further studies have shown that the economic burden of MDD has increased during the 2005-2010 period by 21.5 percent in the US, while in Europe the cost is estimated at 1 percent of Gross Domestic Product [8]. The total cost of MDD in 2010 in 30 European countries was estimated at 91.9 billion euros [9].

Etiology of MDD is attributed to a combination of biological factors, environmental/family stressors, and personal vulnerabilities (i.e., psychoemotional/behavioral traits). Epidemiological studies have identified gender, age, and marital status as key demographic factors affecting the onset of MDD [10]. Genetic factors, early childhood adversity, and premorbid personality characteristics have also been suggested as predisposing MDD factors [11]. Perfectionism, low self-esteem, and maladaptive coping strategies [12] are among the key related personality traits.

A structured clinical interview based on DSM criteria is the standard procedure for depression diagnosis [13]. Quantification of the presence and severity of depressive symptomatology is often aided by rating scales completed by a specially trained mental health professional in the context of the clinical interview. The Hamilton Depression Rating Scale (HDRS or HAM-D), also known as Hamilton Rating Scale for Depression (HRSD), is one of the most popular scales in clinical settings. HAM-D assesses the severity of 17 related symptoms, such as depressed mood, suicidal ideation, insomnia, work and interests, psychomotor retardation, agitation, anxiety, and somatic symptoms [14]. Both HAM-D and DSM clinical criteria have been criticized regarding their reliability [15], [16], as diagnosis of MDD is not as consistent as other common medical conditions [17]. In general, "*there is no blood test*" for depression [18] as the disorder lacks biological gold standards [19].

Even recent classification schemes (e.g., DSM-5) run the risk of confusing normal sadness (e.g., bereavement) with depression, raising the likelihood of false positive diagnoses [4]. Depression assessment is a complex process and diagnosis is associated with a significant degree of uncertainty, given the lack of objective boundaries, and the need to evaluate symptoms within the person's current psychosocial context and past history [20]. Diagnostic accuracy typically improves when results from successive clinical assessments, performed over several months, are taken into account [21]. Importantly, a simple "*symptom checklist*" approach is severely limited and diagnosis requires considerable time investment in order to develop rapport with the patient [18]. The validity and clinical significance of strict classification schemes has also been questioned [22]. For instance, MDD has been questioned as a "*homogeneous categorical entity*" [11], and the notion of a "*continuum of depressive disorders*" is often advocated [23]. These reasonable concerns go beyond the scope of the present review, given that currently affective computing research relies heavily upon established clinical practice tools and procedures.

Clinical diagnosis of depression may also be supported by scores on self-report scales and inventories (Self-RIs). Most often used Self-RIs in affective computing research are the varius forms of PHQ-2/8/9 (Patient Health Questionnaire, comprised of 2, 8, or 9 items, respectively) and Beck's Depression Inventory (BDI); Depression and Somatic Symptoms Scale (DSSS) was also used in one study. Self-RIs are convenient and economical, with reported sensitivity and specificity approaching 80-90 percent (e.g., PHQ-9 [24]), but encompass certain disadvantages. Importantly, they do not take into account the clinical significance of reported symptoms, and do not permit adjustments for individual trait characteristics, other psychiatric and medical comorbidities, and potentially important life events, as opposed to a clinical interview [25]. Moreover, Self-RIs are limited in their capacity to differentiate between depression subtypes [26]. Additionally Self-RIs are vulnerable to intentional (such as norm defiance) or unintentional reporting bias (e.g., subjective, central tendency [i.e., avoiding extreme responses], social desirability, and acquiescence) [27]. In sum, although Self-RIs alone are insufficient to support the diagnosis of depression [28], [29], [30], they are widely used for screening purposes in various settings, including primary health care. While the cost-effectiveness of widespread screening practices for improving the quality of depression care is debated [31], practical issues related to the aforementioned limitations of Self-RIs raise questions regarding the overall utility and effectiveness of this practice for population-based mental health.

Objective measures of psychoemotional state, which are implicitly desirable in clinical and research applications alike [32], [33], could complement Self-RIs and help overcome some of their shortcomings. Certain Self-RIs are sufficiently brief and can be completed on a regular basis (e.g., monthly or weekly) as part of electronic platforms designed to support long-term monitoring of persons at risk. As suggested by Girard and Cohn [34], technological advances in the field have paved the way for viable automated methods for measuring signs of depression with multiple potential clinical applications. Thus, decision support systems capable of capturing and interpreting nonverbal depression-related cues, combined with verbal reports (Self-RIs), could be valuable in both clinical and research applications. In principle, such measures may reduce or even eliminate report bias. In addition, such measures are minimally invasive and do not require extra effort on the part of the respondent, thus likely to increase long-term compliance.

1.2 Investigating Automatic Depression Assessment

Current technological means can provide the infrastructure for continuous monitoring of the psychoemotional state in high-risk individuals as part of early detection and/or relapse prevention programs, such as the SEMEOTICONS[1] EU funded project, aiming to provide reliable indices of anxiety/stress-related cardiometabolic risk [49], [50]. A system devoted to the assessment of depressive symptomatology based on visual cues, should likewise provide reliable indices, partly based on facial expression analysis, in an unobtrusive manner.

Currently, video-based systems for depression assessment have only been found in research-related projects, and have not been applied in the general population to evaluate their feasibility. Although currently limited to research applications, the field has been very popular, with a dedicated section within the "*Audio/Visual Emotion Challenge*" (AVEC). AVEC'13 had three papers accepted for the "*Depression Recognition Sub-challenge*" (DSC) [51], [52], while AVEC'14 [53] respectively attracted 44 submissions by 13 teams worldwide; the latest AVEC (2016), attracted submissions from 7 teams for the DSC [54]. Apart from being an active field drawing broad interest, AVEC submissions document the sheer number of research groups working towards the development of such methods. This fact implies that the idea of developing automated depression assessment methods is not only promising, but is continuously progressing towards more robust and reliable measures. Furthermore, the widespread and relatively low-cost accessibility to computer and internet technologies, webcams, and smart phones, renders an efficient system for depression assessment viable. Practical issues involved in developing such a system, like the storage of sensitive data, could become an issue if not handled properly, but there are ways to tackle such challenges; encryption, protection by password, or even an authorization procedure could be implemented to regulate access to sensitive personal data.

In parallel, a great number of Web-based tools for depression management have been developed and used clinically displaying a high degree of acceptance and patient adherence [55]. However, as it will become apparent in subsequent sections, given the current state of-the-art, video-based systems for depression assessment are not intended as standalone tools, but mainly to function as a decision support systems assisting mental health professionals in the monitoring of persons at risk. "Behaviormedics" is the term Valstar [56] introduced for applications designed for the automatic analysis of affective and social signals, among others, to support objective diagnosis. Finally, the development of tools to assist clinicians in the diagnosis and monitoring response to treatment and illness progression is gradually being supported by clinical studies [57]. For the purpose of the present review, the term "*depression assessment*" will refer to the process of detecting and assessing the severity of signs of and/or presence of depression.

TABLE 1
Search Terms and Web-Resources Employed in the Current Review

Keywords	Web-resources
• Depression • Facial Expression • Non-verbal communication • Image Processing • machine learning • Biomedical Imaging • Face • Emotion • Computer Vision	• ACM Digital Library [35] • IEEE Xplore Digital Library [36] • Elsevier [37] • Springer [38] • Wiley Online Library [39] • NASA [40] • Oxford University Press [41] • US National Library of Medicine [42] • Scopus [43] • Google Scholar [44] • Medpilot [45]
• Depression • Definition • Types •Frequency or rate • Diagnostic tests • Etiology and risk factors • Predictability	• Mayo Clinic [46] • Survey of Health, Ageing and Retirement in Europe [5] • National Comorbidity Survey [47] • World Health Organization [1] • World Health Organization - Regional Office for Europe [48]

Note: The first row contains keywords and web-resources that were canvassed to identify relevant approaches. Elements pertaining to the clinical relevance of studies are listed in the bottom row.

1.3 Inclusion Criteria

The technical report entitled "*Procedures for Performing Systematic Reviews*" by Kitchenham [58] was used as a guide for the present review. The keywords used to search electronic databases and related resources are listed in Table 1; the keywords were used interchangeably, in combinations of two or more, with either "OR" or "AND" operands. Inclusion criteria for the review involved a) adequate description of an algorithm for automatic depression assessment utilizing visual cues, and b) presentation of systematically derived data, producing concrete results. Strictly clinical studies, as well as approaches solely relying on speech-derived cues, were not included.

Publication dates of reviewed studies meeting criteria range between 2007 and April of 2017 (publications since 2005 were considered). Fig. 1 illustrates the rapid increase of relevant studies during the past few years proving that automatic depression assessment based on visual cues is a rapidly growing research domain. The small drop in 2015 can be attributed to the fact that there was no Depression sub-challenge in the 2015 AVEC; similarly, the sharp rise of interest in 2013, 2014, and 2016 can be attributed to the respective AVEC challenges.

The current review focuses on the following study features: the research question addressed (detection/severity assessment/prediction), the number of modalities employed, facial signs, facial regions of interest (ROIs), number of participants (control/patients), technical specifications (image resolution/frame rate), experimental protocols for dataset

[1] http://www.semeoticons.eu/

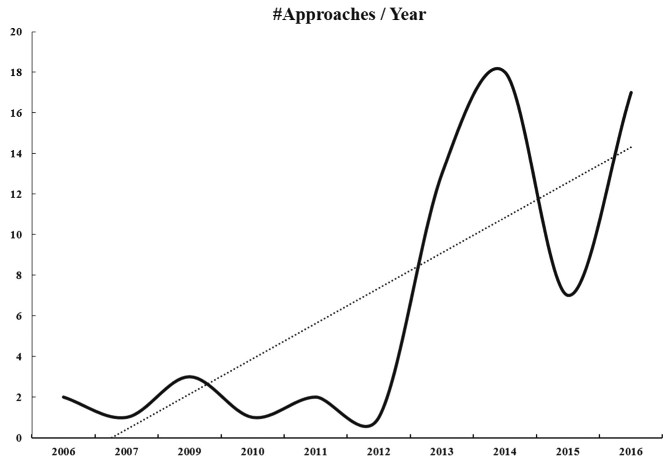

Fig. 1. Number of studies in the field of depression assessment by year of publication.

acquisition, feature descriptors, fusion algorithms, decision methods, and scores.

1.4 Related Work

Despite the rising interest in the topic, existing reviews vary in their specific focus, and rarely attempt an in depth assessment of methods and results. For instance, in their report on health-enabling technologies for addiction and depression detection Jähne-raden et al. [59] reviewed three studies addressing depression detection based on facial activity. Valstar [56] included a non-technical description of five relevant studies in his work on "Automatic Behavior Understanding in Medicine", under the mood and anxiety disorders section. D'Mello [60] reviewed five publications relevant to depression assessment within the general context of mental state detection. Schuller [61] discussed many topics of shared interest with the present review, as did Martinez and Valstar [62], but within the broader field of affective computing and facial image analysis, and not specifically focusing on depression assessment.

Hyett et al. [63] reviewed approaches for the detection of melancholia, including eight papers related to depression, and went into depth analyzing the utility of a chosen set of algorithms. The most extensive review to date by Girard and Cohn [34] includes twenty-one studies, presenting the core concepts and providing the necessary information for someone who is getting acquainted with the field, without, in our view, expanding on many technical details. Cummins et al. [64] conducted an extensive review of depression and suicide risk assessment, with a focus on speech analysis. Finally, the survey by Corneanu et al. [65] is adequately thorough with respect to the algorithms used for facial expression recognition, but is limited to only ten applications to depression assessment.

In the current exhaustive review of more than sixty studies, technical details, potential limitations of each approach and classification accuracy achieved are evaluated, focusing on image processing and machine learning algorithms applied to depression detection. Additional modalities employed (speech, physiological signals, contextual information) and computational cost factors related to system requirements, e.g., camera resolution and frame rate, are also considered.

1.5 Structure of the Paper

The current review is organized in seven sections. Section 2 covers nonverbal assessment of depression and summarizes the visual signs identified in the reviewed studies. The relevant datasets used for evaluating systems for automatic depression assessment are described in Section 3, along with respective data collection procedures. Section 4 reviews image processing and machine learning algorithms used for automatic assessment of depression, while Section 5 presents a quantitative meta-analysis of selected studies. Discussion of the main review findings and implications for future studies are included in Section 6.

2 Nonverbal Signs for Depression Assessment

It is well known that depression manifests through a variety of nonverbal signs [66], [67]. Involuntary changes in the tonic activity of facial muscles, as well as changes in peripheral blood pressure, and skin electrodermal response, often mirror the frequent and persistent negative thoughts and feelings of sadness that characterize depression. Preliminary findings suggest that electroencephalographic recordings may contain features related to depression [68]. Functional Near-Infrared Spectroscopy (fNIRS) has also attracted interest [69], [70]. Additionally, speech conveys non-verbal information on the mental state of the speaker; prosodic, source, and acoustic features, as well as vocal tract dynamics, are speech-related features affected by depression [64]. Furthermore, depression as a mood disorder, is portrayed on the individual's appearance, in terms of facial expression, as well as body posture [66], [67]. Face as a whole, and individual facial features, such as eyes, eyebrows or mouth, are of particular interest when it comes to depression assessment. Some of the visual signs identified in the reviewed papers are briefly described in the paragraphs that follow.

A visual sign that has drawn considerable attention by clinicians in relation to depression assessment is pupil dilation. Siegle et al. [71] reported faster pupillary responses in non-depressed individuals to positive rather than negative stimuli. In contrast, depressed persons displayed slower pupil dilation responses to positive stimuli in conditions associated with reduced cognitive load (see also [72], [73], [74], [75], [76], [77]). More recently Price et al. [78] investigated attentional bias, including pupil bias and diameter, to predict depression symptoms over a two year follow up period in a sample of adolescents displaying high ratings of anxiety. Saccadic eye movements have also been found to differ in terms of latency and duration between depressed and healthy participants [79], [80].

Action Units (AUs), introduced by Ekman et al. [81], were first utilized for automatic depression assessment by Cohn et al. [82]. AUs have been studied in terms of frequency of occurrence, mean duration, onset/total duration, and onset/offset ratios. At approximately the same time McIntyre et al. [83] proposed an AU-based approach in the form of Region Units (RUs). Several studies have reported promising results on the application of AUs to automatic depression assessment [84], [85], [86], [87], [88], [89], [90], [91], [92], [93], [94], [95], [96].

TABLE 2
Non-Verbal Manifestations of Depression

Pupil dilation/bias
Pupillary response
Iris movement
Eyelid activity (openings, blinking)
Saccadic eye movements
Eye gaze (limited & shorter eye contact)
Visual fixation
Low frequency & duration of glances
Extended activity on the corrugator[1] muscle
Eyebrow activity
"Veraguth fold"[2]
Frowns
Fewer smiles
More frequent lip presses
Smile intensity & duration
Mouth corners angled down
Mouth animation
Listening smiles (smiling while not speaking)
Reduced activity on the zygomaticus[3]
Facial activity
Action Units occurrence (mean duration, ratio of {onset/total duration, onset/offset})
Region Units (RUs)
Facial expression occurrence (variability & intensity)
Sad/negative/neutral expression occurrence
Head pose (orientation, movement)
Body gestures (full or upper body, or body parts)
Slumped posture
Limp & uniform body posture
Reduced & slowed arm and hand movements
Shaking and/or fidgeting behavior
Self-adaptors
Foot tapping
Motor variability

[1] *Corrugator is the muscle close to the eye, in the medial extremity of the eyebrow.*
[2] *"veraguth fold" is a fold (wrinkle) of skin on the upper eyelid, between the eyebrows.*
[3] *zygomaticus is the muscle which draws the angle of the mouth to produce a smile.*

Specific facial expressions, have also been examined for depression assessment, in terms of frequency of occurrence, variability, and intensity of a specific expression. Typically, the facial expression classification system proposed by Ekman [97] is employed, which includes a set of six basic emotions (joy, surprise, anger, sadness, disgust, fear). Measuring the frequency of occurrence of each of the six emotional expressions [75], [84], [85], [86], [98], [99], [100] relies on the premise that depressed individuals tend to show reduced expressivity [66]. Other studies focused on specific facial features, such as the eyes and mouth. These include gaze direction [100], [101], [102], [103], reduced eye contact [104], eyelid activity [105], eye openings/blinking [75], [106], [107] and iris movement [106]. Smile intensity [100], [101], [103], smile duration [101], [103], mouth animation [107], listening smiles (smiles while not speaking) [102], and lack of smiles [104] also constitute potentially useful facial signs for assessing depression.

Head pose, orientation, and movement have been used extensively for depression assessment [75], [84], [85], [86], [90], [98], [99], [100], [101], [103], [106], [107], [108], [109], [110], [111], along with motor variability and general facial animation [84], [86], [98], [107]. Additionally, body gestures [109], [110], [111], [112], [113] involving the entire body, the upper body, or separate body parts can also contribute to the assessment. Finally, shaking and/or fidgeting behavior, self-adaptors, and foot tapping [102] have also been considered as signs of depression. Table 2, inspired by [64], [98], [103], [114] and enhanced hereby, is summarizing all of the signs and signals related to depression assessment as found in the literature to date.

3 Depression Datasets

The availability of empirical data is of paramount importance for the evaluation of methods for automatic assessment of depression. Such data are critical during algorithm development and testing. Due to the sensitive nature of clinical data, availability is neither wide nor free, and this is the reason that most research groups resort to generating their own data sets. This section describes procedures used for data collection and derived datasets found in the reviewed studies (cf. Table 3).

3.1 Data Collection Procedure

Recruitment of participants is perhaps the most challenging step in this line of research. Patients with MDD were recruited from the community, in many cases by clinical psychologists or social workers, and were assessed using DSM-IV [115] criteria [71], [72], [74], [82], [90], [111] and/or HAM-D scores [82], [91], [111], [113]; patients may be medicated, un-medicated or in remission. The Mini International Neuropsychiatric Interview (MINI) was employed in the data collection for the dataset reported in [116] in order to obtain the diagnosis, and Quick Inventory of Depressive Symptomatology-Self Report (QIDS-SR) for defining the severity. BDI has also been used in [71] to establish whether a given patient was in remission. Comparison data were obtained from individuals who had never been diagnosed with depression or other mood disorder. Data collection from non-clinical samples, employed Self-RIs such as PHQ-9 [84], [85], [86], [98], [101], [102] and BDI [51], [53], assessing the severity of (sub-clinical) depression-related symptomatology. Recruitment methods further included flyers, posters, institutional mailing lists, social networks, and personal contacts.

In order to ensure that the collected data carry useful information, the experimental protocol must be carefully designed. Information on participant characteristics includes cognitive abilities, assessed mainly through executive function tests, and psychoemotional traits—assessed through self-report questionnaires. Across studies, executive tasks include sorting [71], planning, and problem solving tasks [117]. For instance, the dataset constructed for the AVEC'13 integrated a series of "activation tasks", including vowel pronunciation, solving a task out loud, counting from 1 to 10, reading novel excerpts, singing, and describing a specific scene displayed in pictorial form [51].

Establishing conditions which enable the collection of signs related to depression is by far the most important step, as also discussed in [64]. Methods employed can vary significantly across studies. Emotion elicitation is used to measure reactions to emotionally charged stimuli, given

TABLE 3
Datasets Employed by the Reviewed Studies for Depression Assessment

Corpus	Population / Total (Control / Study)	Symptomatology Collection Methods	Ground Truth	Selection Criteria	Research Question	Image Resolution / Frame Rate	Availability to Third Parties
Pittsburgh	Adults / 49 (-/-)	Interpersonal	Clinical Assessment	DSM-IV, HAM-D > 15	Detection	640x480 / 29.97	Visual & Audio Features (Expected)
BlackDog	Adults / 130 (70/60)	Combination	Clinical Assessment	Control: No history of mental illness, Study: DSM-IV, HAM-D > 15	Detection	800x600 / 24.94	-
DAIC-WOZ	Adults / 189 (-/-)	Combination	Self-report	Age, language, eye-sight	Detection, Severity	-	Visual & Audio Features, Audio Recordings, Transcripts
AVEC	Adults 58 (-/-)	Non-social	Self-report	-	Severity	-	Full Video Recordings, Visual & Audio Features
ORI	Adolescents / 8 (4/4)	Interpersonal	-	-	Detection	-	-
ORYGEN	Adolescents / 30 (15/15)	Interpersonal	Clinical Assessment	Stage1: No depression, age 9-12 years Stage2: Depression, 2 years after	Prediction	-	-
CHI-MEI	Adults / 26 (13/13)	Combination	Clinical Assessment	DSSS, HAM-D	Unipolar Depression / Bipolar Disorder	640x480 / 30	-
EMORY	Adults / 7 (-/7)	Interpersonal	Clinical Assessment	DBS-SCC Treatment	Recovery	-/30	-

that such reactions significantly differ between healthy and patient groups. The Handbook of Emotion Elicitation and Assessment [118] describes several methods for eliciting emotion in the laboratory including: emotional film clips used in [83], [109], images selected from the International Affective Picture System (IAPS) used in [83], [109], social psychological methods, and dyadic interaction. Emotionally charged images and clips are in principle capable of eliciting observable responses, although ethical considerations set limits to the shocking nature of the content. Social experiments also raise certain ethical issues and may not be suitable for emotionally vulnerable persons. In this regard it is imperative that patients with depression are not subjected to unnecessary and unwanted stress or anxiety.

Dyadic interaction is an appealing method as it involves social context, and affords a wide range of elicited emotions [64]. Contemporary models of emotional expression propose that the intensity of relevant signs is proportional to the degree of sociality of the eliciting situation, ranging from being physically in the same room with another person, communicating over the phone, to simply thinking of the other person [119], [120]. The suitability of dyadic interaction contexts is further supported by the notion that many of the behavioral indicators of depression are asocial in nature [34]. A shortcoming of this method lies in its unstructured nature, which does not guarantee that the targeted social or emotional responses will actually be produced.

Structured Interviews are usually employed for gathering depressive symptoms, but have also been used for eliciting specific emotions by asking participants to describe personal, emotionally charged events [83]. Interviews can take place over one or more sessions, conducted by a therapist or a virtual character (VC), or guided by instructions on a computer screen. Typically the interview topic changes smoothly from casual to more intimate topics [82], [83], [84], [86], [87], [90], [91], [98], [99], [101], [103], [109], [117].

The amount of visual data that is necessary for a reliable assessment depends heavily upon the temporal nature of MDD. The specificity of the assessment method may benefit from multiple recording sessions, such as that of the data reported in [82], [90], [91]. Recording length depends on the elicitation method, with structured interviews being considerably longer in comparison with recordings based on emotion elicitation through films.

Additional modalities, such as speech [82], [85], [93], [98], [99], [100], [101] [102], [103], [107], [109], [110], [111], [121], [122], [123], [124], [125], [126], [127], [128], [129], physiological signals [102], [130], [131], and written text [85], [99] have been employed in order to enhance assessment sensitivity. Consequently, studies vary widely depending on the types of equipment utilized, the different modalities and particular signs monitored. For instance, studies focusing on the pupillary response may only use a pupilometer [71], [72], [73], [74] and pay special attention to ambient illumination in order to optimize sensitivity. Again, depending on the approach, one or more cameras, typically color, are simultaneously used to cover more than one viewing angles and fields of view (e.g., both face and body separately [82]). Thermal imaging has

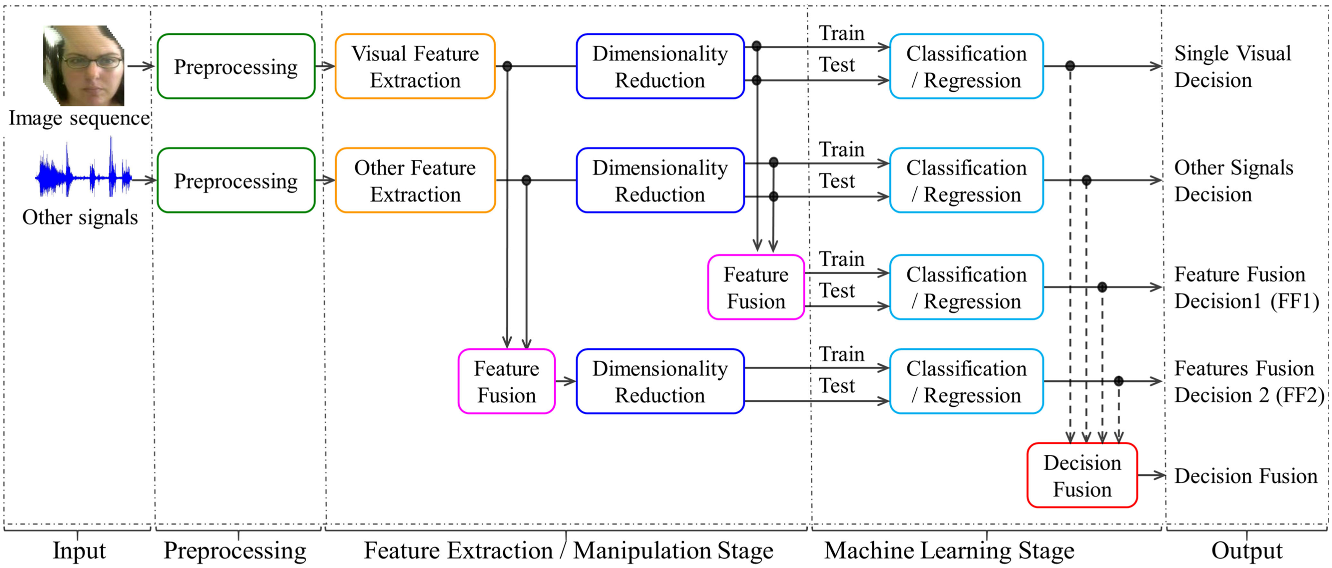

Fig. 2. Workflow for automatic depression assessment. The output can be derived from a) single feature sets/modalities, b) feature fusion, with dimensionality reduction before (FF1) or after fusion (FF2), and c) decision fusion with any possible combination of outputs from single feature sets/modalities and feature fusion.

been used to investigate clinical severity of depression based on eye temperature[132], as well as in studies focusing on saccadic eye movements, employing an infrared eye tracking system [79], [80]. Depth sensors (e.g., Microsoft Kinect) have also been utilized in some cases [84], [101]. Microphones are naturally required for recording the participants' speech during an interview or narrating tasks. Again, as with emotion elicitation, the precise setup varies across studies.

3.2 Reported Datasets

The various datasets reported in relevant work are summarized in Table 3. Participant demographics, stimuli, ground truth, selection criteria, research question, as well as technical specifications, are some of the features that vary across studies. Most of the studies employed adult participants, while two recruited adolescents. Methods for collecting depressive symptomatology included: a) interpersonal, i.e., interview with a clinical psychologist or interaction with family members, b) non-social, where participants were presented with stimuli on a computer, and c) combination of (a) and (b). The ground truth for the presence of depression varied accordingly, relying on clinical assessment in the majority of the cases, and on self-reports in two of the studies.

The selection criteria used depended greatly on the research question. DSM and HAM-D criteria were used for detection of depression [82], [90], [110], [111], [113], [132] or differentiation from Bipolar Disorder [95]. Others had more specific criteria, i.e., patients recovering from Deep Brain Stimulation of the Subcallosal Cingulate Cortex (DBS-SCC) [133], in order to monitor recovery progress. Studies assessing the predictive value of the method for future emergence of clinical depression in adolescents involved 9-12 year old participants at the initial data collection, with clinical reassessment after a two year interval [117], [125].

Technical specifications of the video recording equipment varied to some extent, but not significantly, as the setup typically involved a single camera monitoring the participants' face/upper body. A notable exception was the setup employed for the Pittsburgh dataset, utilizing four hardware-synchronized analogue cameras; two for monitoring the participant's head and shoulders, one for full body monitoring, and the fourth monitoring the interviewer, together with two microphones for speech recording.

Regarding dataset availability, AVEC is the only fully available dataset for free download,[2] while the DAIC-WOZ dataset (Distress Analysis Interview Corpus—Wizard of Oz) is also partly available.[3] Both datasets require a signed End User License Agreement (EULA) in order to provide download access. Pittsburgh dataset is also expected to release features by June 2017.[4] The remaining reported datasets are proprietary, while in some cases they have been made available to visiting researchers. The number of participants listed in Table 3 is that reported in the latest publication employing the related dataset. However, different published papers report results obtained from different subsets; accordingly, sample size used in each published report is specified in Section 5 (Tables 7 and 8).

4 Automatic Assessment of Depression

The generic processing flow of an automatic system for depression assessment, combining the standard structure for automated audiovisual behavior analysis proposed by Girard and Cohn [34], with fusion methods presented in Alghowinem's thesis [134], is illustrated in Fig. 2. Given a visual input (image sequence), along with other types of signals, such as audio, text from transcripts, and physiological signals, the prerequisite step is that of preprocessing. Feature extraction algorithms are subsequently applied to all visual signals, as described in Section 4.2.1 and illustrated in Fig. 3, while methods for dimensionality reduction and feature level fusion are reported in Sections 4.2.2 and 4.2.3, respectively. Machine learning algorithms are employed, depending on the research question, i.e., presence of depression or severity assessment. Classification approaches are

[2] http://avec2013-db.sspnet.eu/
[3] http://dcapswoz.ict.usc.edu/
[4] http://www.pitt.edu/ emotion/depression.html

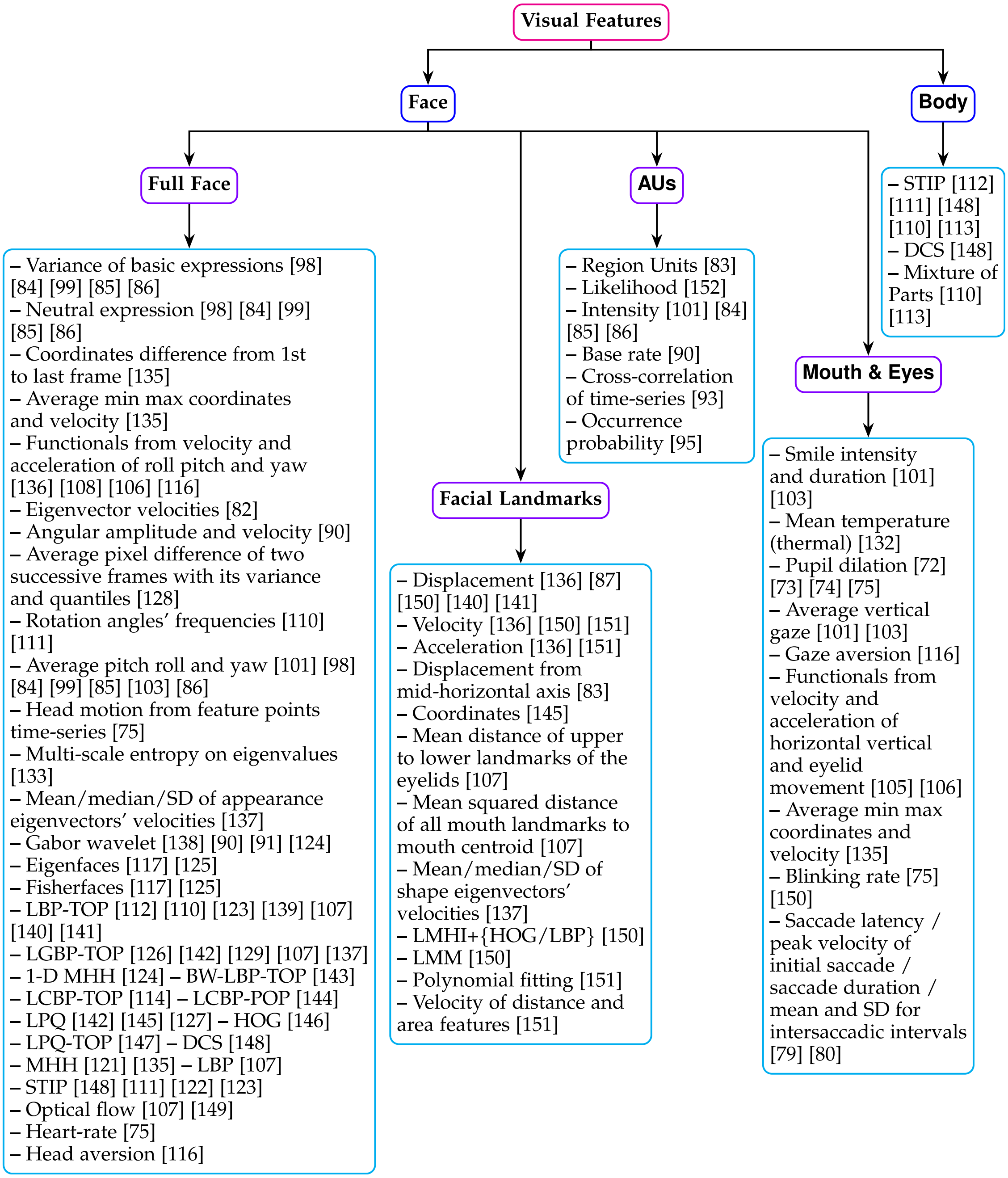
Visual Features
Face
Body
Full Face
AUs
Facial Landmarks
Mouth & Eyes
– Variance of basic expressions [98] [84] [99] [85] [86]
– Neutral expression [98] [84] [99] [85] [86]
– Coordinates difference from 1st to last frame [135]
– Average min max coordinates and velocity [135]
– Functionals from velocity and acceleration of roll pitch and yaw [136] [108] [106] [116]
– Eigenvector velocities [82]
– Angular amplitude and velocity [90]
– Average pixel difference of two successive frames with its variance and quantiles [128]
– Rotation angles' frequencies [110] [111]
– Average pitch roll and yaw [101] [98] [84] [99] [85] [103] [86]
– Head motion from feature points time-series [75]
– Multi-scale entropy on eigenvalues [133]
– Mean/median/SD of appearance eigenvectors' velocities [137]
– Gabor wavelet [138] [90] [91] [124]
– Eigenfaces [117] [125]
– Fisherfaces [117] [125]
– LBP-TOP [112] [110] [123] [139] [107] [140] [141]
– LGBP-TOP [126] [142] [129] [107] [137]
– 1-D MHH [124] – BW-LBP-TOP [143]
– LCBP-TOP [114] – LCBP-POP [144]
– LPQ [142] [145] [127] – HOG [146]
– LPQ-TOP [147] – DCS [148]
– MHH [121] [135] – LBP [107]
– STIP [148] [111] [122] [123]
– Optical flow [107] [149]
– Heart-rate [75]
– Head aversion [116]
– Region Units [83]
– Likelihood [152]
– Intensity [101] [84] [85] [86]
– Base rate [90]
– Cross-correlation of time-series [93]
– Occurrence probability [95]
– Displacement [136] [87] [150] [140] [141]
– Velocity [136] [150] [151]
– Acceleration [136] [151]
– Displacement from mid-horizontal axis [83]
– Coordinates [145]
– Mean distance of upper to lower landmarks of the eyelids [107]
– Mean squared distance of all mouth landmarks to mouth centroid [107]
– Mean/median/SD of shape eigenvectors' velocities [137]
– LMHI+{HOG/LBP} [150]
– LMM [150]
– Polynomial fitting [151]
– Velocity of distance and area features [151]
– STIP [112] [111] [148] [110] [113]
– DCS [148]
– Mixture of Parts [110] [113]
– Smile intensity and duration [101] [103]
– Mean temperature (thermal) [132]
– Pupil dilation [72] [73] [74] [75]
– Average vertical gaze [101] [103]
– Gaze aversion [116]
– Functionals from velocity and acceleration of horizontal vertical and eyelid movement [105] [106]
– Average min max coordinates and velocity [135]
– Blinking rate [75] [150]
– Saccade latency / peak velocity of initial saccade / saccade duration / mean and SD for intersaccadic intervals [79] [80]

Fig. 3. Taxonomy of visual features utilized in the reviewed studies for depression assessment.

suitable for categorical assessment, such as discriminating between a given number of classes (i.e., depressed versus not depressed or low versus high depression severity). For continuous assessment of depression (e.g., depression severity according to BDI scores ranging between 0-63) regression analyses are more appropriate. Fusion from different feature sets, either visual or from different modalities, can also take place at the decision phase (more details are given in Section 4.3.4). Methods used in the reviewed studies (i.e., algorithms for feature extraction from visual signs, feature selection algorithms, decision methods, fusion techniques) are described in turn in the following section. Given that the present systematic review is focused on depression assessment, a more extensive description of relevant methods can be found in Corneanu et al. [65] with respect to algorithms for facial expression recognition in generally.

4.1 Preprocessing

Given a visual input (video), illumination normalization, registration and alignment between the image sequences, and face detection are typical required preprocessing steps. Other types of signals, such as speech or physiological recordings, may also need preprocessing, such as segmentation. The most popular algorithm for face detection has been proposed by Viola and Jones [153]. Some off-the-shelf facial expression analysis applications have also been used widely as preprocessing tools, enabling researchers to focus on deriving high level information. An example of such a tool is the OpenFace freeware application[5] [154]. SEMAINE API, an open source framework for building emotion-oriented systems, is another potentially useful preprocessing tool

[5] https://www.cl.cam.ac.uk/ tb346/res/openface.html

TABLE 4
Features Associated with Statistically Significant Differences Between Study and Control Groups, as Reported in the Original Papers (Similar to Cummins et al. [64])

Feature	Reference	Study Group	Population (Control/Study) / Male rate	Control (mean ± S.D.)	Study (mean ± S.D.)	Significance
Saccade latency	Winograd-Gurvich et al. (2006) [80]	Melancholic depression	24 (15/9)/25%	151.3 ± 16.7	172.0 ± 22.1	$p < 0.05$ (post hoc)
Variability of saccade latency	Winograd-Gurvich et al. (2006) [80]	Melancholic depression	24 (15/9)/25%	24.4 ± 77.0	48.1 ± 35.7	$p < 0.05$ (post hoc)
Average base rate (ABR) of AU14	Girard et al. (2013) [90]	High vs low depression severity	19 (-/-)/36.8%	17.0%	27.8%	$p < 0.05$ (Wilcoxon signed rank test)
ABR of AU15	Girard et al. (2013) [90]	High vs low depression severity	19 (-/-)/36.8%	16.5%	8.5%	$p < 0.05$ (Wilcoxon Signed Rank test)
Average mean square of head motion (AMSHM) vertical amplitude	Girard et al. (2013) [90]	High vs low depression severity	19 (-/-)/36.8%	0.0029	0.0013	$p < 0.05$ (Wilcoxon signed rank test)
AMSHM vertical velocity	Girard et al. (2013) [90]	High vs low depression severity	19 (-/-)/36.8%	0.0005	0.0001	$p < 0.01$ (Wilcoxon signed rank test)
AMSHM horizontal amplitude	Girard et al. (2013) [90]	High vs low depression severity	19 (-/-)/36.8%	0.0034	0.0014	$p < 0.01$ (Wilcoxon Signed Rank test)
AMSHM horizontal velocity	Girard et al. (2013) [90]	High vs low depression severity	19 (-/-)/36.8%	0.0005	0.0002	$p < 0.01$ (Wilcoxon signed rank test)
Pupil area in darkness	Wang et al. (2014) [74]	Depression	30 (15/15)/0%	26.0 ± 5.1	32.4 ± 5.4	$p < 0.01$ (t-test)
Pupil area in largest constriction	Wang et al. (2014) [74]	Depression	30 (15/15)/0%	16.9 ± 3.8	20.8 ± 5.8	$p < 0.05$ (t-test)
Smile intensity	Scherer et al. (2014) [98]	Depression	111 (79/32)/-	19.9 ± 16.9	12.8 ± 11.1	$p < 0.05$ (t-test)

[155]. The Computer Expression Recognition Toolbox (CERT) [156] has been quite popular, but has now become commercialized. Tools for gaze estimation, such as the one presented in [157], [158], may help derive important features relevant to signs of depression, such as fixation or shorter eye-contact. Z-Face [159] has also been employed for alignment and landmark detection.

4.2 Feature Extraction / Manipulation

This section describes processes involved in feature extraction, dimensionality reduction, and fusion. The output of this processing stage generates the input to the machine learning stage, where no further manipulation of features is taking place.

4.2.1 Feature Extraction

Feature extraction is an important step in the processing workflow, since subsequent steps entirely depend on it. The approaches reviewed employ a wide range of feature extraction algorithms which, according to the well-established taxonomy in [65], can be classified as a) geometry-based, or b) appearance-based. In the field of depression assessment, several features are derived from the time-series of both (a) and (b) in the form of dynamic features. Close inspection of depression manifestations, listed in Table 2, reveals that the majority of signs involve muscle activity, which accounts for the temporal nature of the features. Features can be further categorized as high or low level; high level features directly translate to human common sense, while low level features are based on "traditional" image processing descriptors. Depending on the approach, the software packages mentioned in Preprocessing could also serve as feature extraction methods (e.g., OpenFace, SEMAINE API, CERT, etc). In the present paper feature extraction algorithms are grouped into those focusing on the face region and those relying on the body region. A pictorial taxonomy of the various algorithms, including the region of interest on which they are applied, the features computed, and references to respective studies, is presented in Fig. 3. Features associated with statistically significant differences between study and control groups, as reported in the original papers, are presented in Table 4. The various features, retrieved from relevant studies, are described below in detail.

Face

Features related to the face are classified here into features from full face, AUs, facial landmarks, and mouth/eyes.

Full Face. As it becomes apparent from Fig. 3, approaches employing feature extraction from the entire face region comprise by far the most popular category. Certain high level features extracted from the face as a whole concern basic emotional expressions displayed, given that depression is associated with reduced variability of emotional expression and greater persistence of a neutral expression. Heart-rate, derived unobtrusively from facial images, has also been used as a feature for detecting depression.

As expected, geometrical features, such as edges, corners, coordinates, and orientation, are often used to represent facial expressions. Functionals derived from the time series of geometric features are quite popular. Some examples are average, minimum, and maximum values of displacements, velocities, or accelerations of the coordinates that define the face region as a whole. Functionals from roll, pitch, and yaw, have also been employed in some approaches, along with the frequency of certain rotation angles. Other approaches go one step further, to compute functionals from the time series of feature points and eigenvectors, rather than relying on simple coordinates. Eigenvalues have also been used for the computation of multi-scale entropy. In one case the difference of face co-ordinates between the first and the last frame was considered.

Appearance-based algorithms are also very popular for full-face based features. Among the most prevalent texture descriptors are Local Binary Patterns (LBP). For LBP to be computed, the image is divided into partially overlapping windows and each pixel of the window is compared to its neighbors producing a binary value; histograms are then constructed on the occurrences of these values, and all histograms are concatenated to form the feature vector. Several variants of LBPs have been created for automatic depression assessment, such as an extension of LBP that considers patterns on Three Orthogonal Planes (LBP-TOP): XY, XT and YT; XY represents each frame, while XT and YT encode space-time information. Local Gabor Binary Patterns-Three Orthogonal Planes (LGBP-TOP) extends LBP-TOP by computing patterns on the output of Gabor-filtered data, rather than on the original intensity image. Gabor filters have been shown to share many similarities with properties of the human visual system. Along the same lines, Local Curvelet Binary Patterns-Three Orthogonal Planes (LCBP-TOP) was introduced in some studies, which entails computing the patterns on the curvelet transform of the original image. Local Curvelet Binary Patterns- Pairwise Orthogonal Planes (LCBP-POP) is yet another variation of the algorithm operating on pairs of orthogonal planes. Additionally, the Block-Wise LBP-TOP (BW-LBP-TOP) method which computes the LBP-TOP for a specific number of non-overlapping blocks, has also been employed.

Local Phase Quantization (LPQ) is another texture descriptor computed on the quantized Fourier phase in order to produce the binary coding for each pixel; it has been shown that the derived descriptor is immune to moderate blurring. An extension of LPQ is Local Phase Quantization-Three Orthogonal Planes (LPQ-TOP), based on the same concept as LBP-TOP. Eigenfaces, which applies eigenvector decomposition to facial images and face recognition, and Fisherfaces representing faces on a subspace through Linear Discriminant Analysis (LDA), have also been used for automatic depression assessment. Another popular algorithm for motion-based approaches is the histogram of optical flow, which estimates the motion within visual representations, by using vectors for each pixel in order to describe a dense motion field. Divergence-Curl-Shear (DCS) descriptors are also based on optical flow, whereby optical flow vectors are transformed into the motion features of divergence, curl and shear.

The Motion History Histograms (MHH) algorithm, which extends Motion History Images (MHI), has also been found in the related literature. MHI is a grayscale image representing local motion: white pixels for latest motion, darkest gray intensities for the earliest, and lighter gray values for intermediate latencies. MHH extends MHI by considering patterns of movement across frame series. A further extension of MHH is the 1-D MHH, which is computed on the feature vector sequence, instead of the intensity image. The Difference Image is a simplified process, which considers intensity differences between the first and the last frame. A similar approach considers pixel differences for every two successive frames, and also the variance and quantiles of the average pixel differences between every two frames. Finally, the Space-Time Interest Points (STIP) algorithm, which detects local structures characterized by significant intensity variability in both space and time, has also been used in several reported studies.

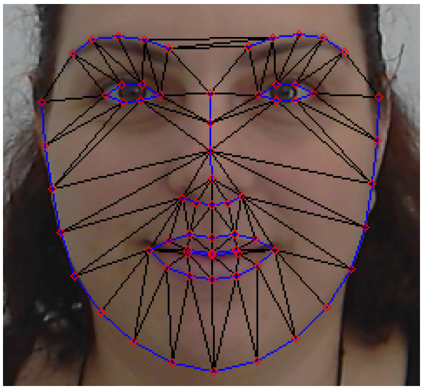

Fig. 4. CLM fitted on the face (created using algorithm from [160]).

Facial Landmarks. Facial landmarks have been very popular in addressing problems related to facial expression analysis, and have been applied to depression assessment. Such algorithms localize fiducial points of the face and facial features, which are very useful in extracting high level traits directly associated with signs of depression, e.g., smiling. The Constraint Local Models method (CLM) introduced by Saragih et al. [160] is displayed in Fig. 4 to illustrate its application to the modeling of facial geometry. Active Shape Models (ASM) as well as Active Appearance Models (AAMs) have also been utilized for depression assessment methods. Other model-based approaches, used in the reported studies, include 3D Landmark Model Matching, Elastic Bunch Graph Matching (EBFM), and Landmark Distribution Model (LDM).

In addition, facial landmark data have also been analyzed as time series. Displacement, velocity, acceleration, as well as the landmark coordinates alone, have been used as features. Furthermore, displacement of each landmark from the mid-horizontal axis, landmark velocity and acceleration have been used as motion-related features. Additionally, velocity vectors of features that represent the relative position of lower-level landmarks have been utilized in some studies, such as the mean distance of upper to lower eyelid landmarks, and mean squared distance of all mouth landmarks to the mouth centroid. In addition, polynomial fitting, as well as statistics derived from shape eigenvector velocities have been utilized. Landmark Motion History Images (LMHI) is a low level feature which, instead of the actual intensities, computes the MHI on the motion of the facial landmarks. LMHI has been combined with Histogram of Oriented Gradients (HOG), as well as with LBP. Finally, the Landmark Motion Magnitude (LMM) algorithm has also been applied to the vectors which displace each landmark from one frame to the next.

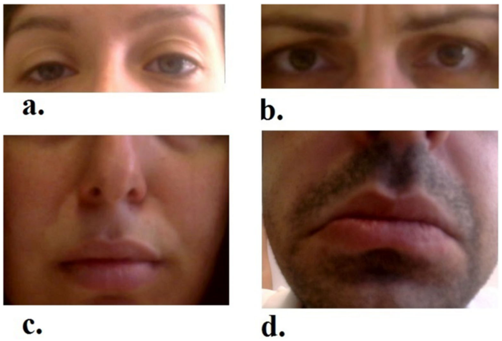

Fig. 5. AU examples typical of depression: a. AU01 inner brow raiser (sadness), b. AU04 brow lowerer (sadness), c. AU11 nasolabial furrow deepener (distress), d. AU15 lip corner depressor (distress).

Action Units. AUs encode the coordinated activity of groups of facial muscles in correspondence to specific actions, including specific emotional expressions. They can be employed for measuring the Variability of Facial Expression (VFE), as depressed individuals tend to be less expressive. Other approaches apply AUs as high-level features. AU occurrence by itself is meaningful as there are specific facial actions that are directly linked to the presence of depression (smiling, mouth corners angled down, etc.). Examples of AUs related to affective states common in depression, such as sadness and distress, according to Emotional Facial Action Coding System (EMFACS) [161] are shown in Fig. 5. It should be clarified, however, that there are AUs which do not necessarily occur as a result of affect-related events, but are associated with non-affective orofacial movements, such as speech and chewing. Additionally, although several approaches implement AUs dynamically (e.g., duration, base rate, ratio of onset/offset), AUs are essentially static signs.

Mouth & Eyes. Apart from the face as a whole, features extracted individually from the mouth and eyes have also been found in the reviewed literature. Smile intensity and duration is a mouth-based feature which has been employed for automatic depression assessment, consistent with the clinical literature, as depressed individuals tend to smile less often.

For the eye region, average vertical gaze, blinking rate, and pupil dilation have been reported. Pupillary response data from pupil radius measurements have been extracted through deformable template matching, which determines the pupil radius by using a pupil model formed by two concentric circles. Additionally, functionals from velocity and acceleration of horizontal and vertical eyelid movement have been used.

Features derived from thermal imaging, such as mean eye temperature, have also been used to differentiate depressed from healthy control samples. Additional features include saccade latency, peak velocity of initial saccade, saccade duration, mean and standard deviation (SD) of intersaccadic intervals.

Body

Although body signs in general have been shown to convey manifestations of depression, few approaches have exploited their utility. Existing applications can be classified as relying on either upper body or relative body part movements. Features for upper body movements have been extracted through the STIP and DCS algorithms. Relative body part movements, on the other hand, have been exploited via the parts algorithm that represents orientation and distance from the torso center expressed in polar coordinates.

4.2.2 Dimensionality Reduction

Many feature extraction algorithms produce vectors of high dimensionality. The goal of dimensionality reduction algorithms is to reduce the number of features in a meaningful manner, in order to avoid corrupting the classifier. Dimensionality reduction algorithms can be classified in two groups: (a) Feature Transformation, and (b) Feature Selection [162]. In the first group features are transformed/combined by being projected from a high-dimensional space to low-dimensional space, to increase separability. On the other hand, in the second group, as the name implies, a selection procedure takes place, and the most discriminative/significant features are selected. Below, examples from both groups, as retrieved in reported approaches, are being described.

Feature Transformation

Principal Components Analysis (PCA) is the most popular algorithm in this category [82], [93], [124], [129], [137], [139], [142], [148], [163] and has been used to generate new features based on a linear transformation of the original features. Manifold learning with Laplacian Eigenmaps [90], [91] supports non-linear dimensionality reduction, based on local computations by solving a sparse eigenvalue problem.

Another set of approaches for reducing dimensionality involves codebooks. Bag-of-Words (BoW) [110], [111], [112], [122], [123], [128], [140], [141], [148], initially intended for document classification, has been applied to image processing problems by treating individual features as words and creating a dictionary of image features. The Vector of Local Aggregated Descriptors (VLAD) [148] also relies on codebook representation. Another dictionary based method is K-SVD (Singular Value Decomposition) [147].

Gaussian Mixture Models (GMM) [105], [108] and Canonical Correlation Analysis (CCA)[145] have been adopted for classification and regression, respectively. In [117], [125] PCA was used in combination with LDA as a feature extraction method, while providing the option of reducing dimensions of the feature vector. A histogram-based approach was presented in [124] which entails maintaining the highest-scoring bins based on a predefined threshold adjusted to the total samples size.

Feature Selection

In [106], [116], and [85] distributional statistics (e.g., t-tests) were employed to select only those features that met a predefined statistical threshold. In [136] the authors implemented the minimum Redundancy Maximum Relevance (mRMR) feature selection, which considers statistical dependency between features.

Several feature selection approaches were evaluated in [107]: supervised feature selection, brute-force selection, and a backward selection scheme using bivariate correlations. Min-Redundancy Max-Relevance (mRMR) [136], [164], greedy forward feature selection [84], [86], relief from WEKA[6] [135], Mutual Information Maximization (MIM)

[6] http://www.cs.waikato.ac.nz/ml/weka/

TABLE 5
Classification Algorithms Employed in the Reviewed Studies

SVM [75], [82], [87], [91], [98], [99], [105], [106], [108], [110], [111], [112], [113], [116], [123], [126], [139]
Nearest Neighbour (NN) [114], [117], [125], [144]
Gaussian Mixture Models (GMM) [105], [108], [138]
Neural Networks (NNet) [111], [112]
Naïve Bayes (NB) [84], [86]
Maximum Entropy Model (MaxEnt) [85], [99]
Random Forest (RF) [146], [150]
Relevance Vector Machines (RVM) [129]
Logistic Regression (LogR) [136]
Hidden Conditional Random Fields (HCRF) [99]
Hidden Markov Model (HMM) [95]
Coupled Hidden Markov Model (CHMM) [95]
Stacked Denoising Autoencoders (SDA) [164]

[151], are additional algorithms used for feature selection in the reviewed studies.

4.2.3 Feature Fusion

Several approaches involve a variety of features derived from different modalities (e.g., visual, audio, text), as well as within the same modality (e.g., visual from different body parts). Fusion methods are usually employed in order to combine multiple feature sets. More often fusion takes place immediately after feature extraction [85], [101], [110], [122], [123], [124], [126], [139], [142], where the extracted feature vectors of the different modalities are concatenated. This concatenation (cf. Fig. 2) can take place before dimensionality reduction (i.e., the concatenated feature set is examined for the dimensionality reduction) or after (i.e., features sets are considered individually for dimensionality reduction).

4.3 Machine Learning

The next step following feature extraction and manipulation, in all methods reviewed, is the machine learning stage. Depending on the particular research goals, different types of decision methods may be applied. Classification methods are appropriate to address categorical questions (e.g., "*depressed*" versus "*non-depressed*" and low versus high depression severity). When the research question concerns the concurrent prediction of depression severity through video-derived indices in a continuous manner, regression approaches are predominantly employed. Cross validation methods are typically applied before classification/regression step.

4.3.1 Cross Validation Methods

Cross validation methods are employed to establish the algorithm reliability, namely its capacity to generalize well with newly introduced data. To establish reliability, a given data set is divided in two parts, one used to train the proposed algorithm, and another (left-out) to test its performance. Specific procedures used for dataset splitting include the leave-one-out [82], [85], [90], [91], [103], [105], [108], [123], [136], [150] and the k-fold method [99], [116], [144], [164]. In the leave-one-out procedure, for a dataset of N samples, N training sets are created of size N-1, each time consisting of all but one sample. The algorithm is then tested N times on its capacity to classify the "*left-out*" cases for each set. Samples could be several for one subject, and therefore the leave-one-out could also be implemented in a leave-one-subject-out manner, where all samples from a specific subject are excluded each time. In the k-fold procedure the dataset is randomly split into k partitions, with one partition kept each time for testing and the remaining used for training the algorithm. This procedure is repeated for k times. In the context of the AVEC challenges, partitioning of the dataset into training and test sections was performed by the organizers, to permit direct comparisons between the algorithms used by participating groups.

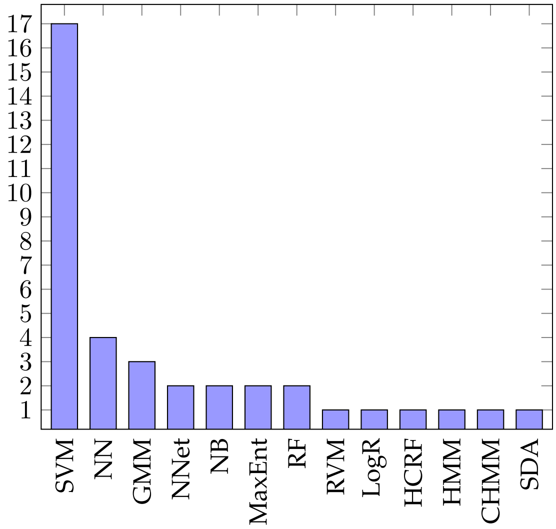

Fig. 6. Ranking of classification algorithms.

4.3.2 Classification

An exhaustive list of classifiers employed in the reviewed studies can be found in Table 5 and were ranked in Fig. 6. Support Vector Machines (SVM) is by far the most popular method for categorical assessment of depression. This can be justified by the fact that SVMs are well suited for binary problems of high dimensionality [165], such as the distinction of low symptom severity/absent depression from high symptom severity/present depression. In addition, SVMs are suitable for non-linear combinations of kernel functions [166]. Furthermore, a variety of Neural Networks have been employed, including Feed Forward Neural Networks, Probabilistic Neural Networks, Restricted Boltzmann Machines (RBM), Radial-basis functions, Multilayer Perceptron, and Extreme Learning Machines.

4.3.3 Regression

The continuous nature of depressive symptomatology is well supported by the clinical literature, as discussed in Section 1.1. As a result, relevant approaches have recently been gaining momentum, including the AVEC challenges aimed at predicting scores on self-report depression scales as a continuous variable using speech and video cues. A common underlying objective in these methods is to develop a function through a combination of features, which can then be used to compute predicted depression severity scores for each participant. As it can be observed in Table 6, the most popular regression algorithm, similarly to the classification-based approaches, is

TABLE 6
Regression Algorithms Employed in the Reviewed Studies

SVR [107], [122], [127], [128], [129], [146], [147], [167]
Linear Regression (LR) [93], [124], [146]
Partial Least Square Regression (PLS) [121], [124]
Canonical Correlation Analysis (CCA) [142], [145]
Gaussian Staircase Regression (GSR) [167], [168]
Discriminative Mapping (DM) [147]
Moore-Penrose Generalized Inverse (MPGI) [142]
Relevance Vector Machines (RVM) [129], [168]
Extreme Learning Machines (ELM) [93]
Random Forest Regressor (RFR) [96]
Deep Convolutional Neural Networks (DCNN) [146]

Support Vector Regression. Again the ranking of algorithms is illustrated in Fig. 7.

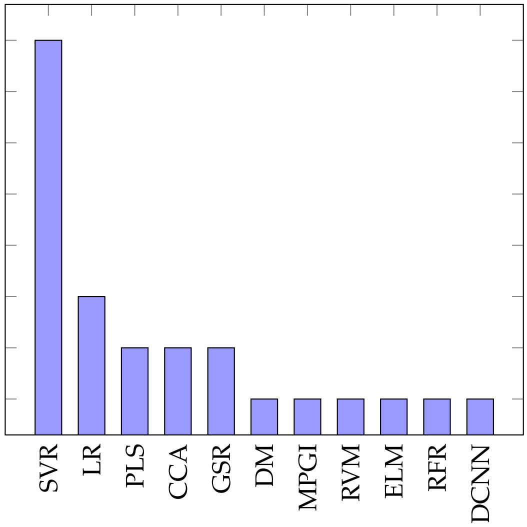

Fig. 7. Ranking of regression algorithms.

4.3.4 Decision Fusion

Fusion at the decision level involves combining different classification/regression results (cf. Fig. 2). The results to be combined can be decisions based on a single modality, but also from feature level fusion, in every possible combination. Fusion for regression-based algorithms is regarded as more challenging due to the multicollinearity which characterizes observational datasets, such as the data used for depression assessment.

Logical operations are the simplest way to implement decision fusion for binary classification (such as for high/low depression severity or presence/absence of depression). AND and OR operators have often been used to combine classification decisions [96], [110], [113], [123], [150]. Training an SVM classifier on scores representing distance from the SVM classifier is another often used approach for decision fusion [110], [123]. Another approach considered SVR output from individual modalities as input to a next-level regressor [146].

Generalized Linear Models [129] and Expectation Maximization algorithms [85] have been applied to decision-level fusion as special cases of regression algorithms. Kalman filtering has also been used to predict the most likely decision [127]. In [121] decision fusion was implemented linearly through a weighted sum according to the formula

$$D_{linear}(\hat{x}) = \sum_{i=1}^{K} w_i D_i(\hat{x}). \tag{1}$$

Where $\hat{x}$ is the test sample, $D_i(\hat{x})$ is the ith decision, and w_i the corresponding weight ($\sum_{i=1}^{K} w_i = 1$).

Williamson et al. [93], [167] also computed a weighted sum according to

$$w_i = R_i^2/(1/R_i^2), \tag{2}$$

where R is the correlation between predicted and actual BDI scores.

5 Selected Approaches & Meta-Analysis

In this section different approaches, either classification- or regression-based, are compared in a quantitative manner. The aim is to derive meaningful conclusions regarding the state of the field, and provide a means to identify the most appropriate setups, both in terms of data collection and algorithms employed. To be included in the analysis, studies must have reported results on automatic assessment of depression using visual features. Deciding on what is the 'best' approach for potential clinical use is beyond the scope of this systematic review. 'Winning' approaches can only be declared in challenges, where conditions are comparable and strictly defined. This is not the case for the various approaches included in this analysis and summarized in Tables 7, 8, and 9, since they were typically evaluated on different datasets (or subsets from the same corpus of data). Even in the case of AVEC participations, direct comparisons across the three challenges are not possible given that different data sets were used in each. The specific objective of our quantitative meta-analysis is to identify general trends, key and strong points, to be considered in future studies of automatic depression assessment, given that a direct comparison of results is not viable.

5.1 Categorical Depression Assessment

Approaches for categorical depression assessment presented in this section are grouped and compared in terms of the employed dataset, in accordance with Table 3. Further, the results are considered with regard to the evaluated features, in reference to the taxonomy presented in Fig. 3. The various approaches, apart from reporting different performance metrics, were tested on datasets or particular subsets of varying sizes. Performance metrics in each report are explained next.

Accuracy, which was reported in the majority of studies, is computed according to Equation (4) based on the following confusion matrix:

$$C = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}, \tag{3}$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives. Certain studies report "*depressed accuracy*", which implies sensitivity (or recall) given by (6)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4}$$

TABLE 7
Comparison of Approaches for Categorical Assessment of Depression Grouped According to the Dataset Used, Ranked within Group Based on Kappa

Data	Paper	Population (Study/Control) / Male rate	Features	Classification Algorithm	Reported Accuracy (or as otherwise noted)	Kappa
Pittsburgh	Joshi (2013) [113] [110]	36 (18/18) / 36.1%	Body, Full Face	SVM	97.2%	0.94
	Alghowinem et al. (2015) [106]	38 (19/19) / 36.8%	Eyes, Full Face	SVM	mean recall =94.7%	0.89
	Dibeklioğlu et al., (2015) [136]	95 (58/37) [*1] / 40.4%	Facial Landmarks, Full Face, Audio	Logistic Regression	91.38%	0.78
	Dibeklioğlu et al., (2017) [164]	130 ([58/35]/37) [*1] [*2] / 40.4%	Facial Landmarks, Full Face, Audio	SDA	78.67%	0.73
	Dibeklioğlu et al., (2015) [136]	-//-	Facial Landmarks	Logistic Regression	84.49%	0.63
	Dibeklioğlu et al., (2017) [164]	130 ([58/35]/37) [*1] [*2] / 40.4%	Facial Landmarks	SDA	72.59%	0.62
	Dibeklioğlu et al., (2015) [136]	-//-	Full Face	Logistic Regression	86.21%	0.60
	Cohn et al. (2009) [82]	107 (66/41) [*1] / 35%	Facial Landmarks	SVM	79%	0.53
BlackDog	Joshi (2013) [110] [123]	60 (30/30) / 50%	Upper Body, Full Face, Audio	SVM	91.7%	0.83
	Alghowinem et al. (2016) [116]	-//-	Full Face, Audio	SVM	mean recall =88.3%	0.77
	Alghowinem et al. (2016) [116]	-//-	Eyes, Full Face	SVM	mean recall =78.3%	0.57
	Alghowinem et al. (2013) [108]	-//-	Full Face	SVM	mean recall =76.8%	0.53
	Joshi 2013, Joshi et al. (2013) [110] [111]	-//-	Upper Body	SVM	76.7%	0.53
	Alghowinem et al. (2015) [106]	-//-	Eyes, Full Face	SVM	mean recall =76.7%	0.53
	Alghowinem et al. (2013) [105]	-//-	Eyes	SVM	mean recall =75%	0.50
DAIC-WOZ	Yang et al. (2016) [146]	35 (7/28) / 45.7%	Facial Landmarks, AUs, Full Face, Audio, Text	Random Forest	F1 = 0.86	0.82
	Scherer et al. (2013) [98]	39 (14/25) / 62%	Full Face, Audio	SVM	89.74%	0.76 [*3]
	Yu et al. (2013) [99]	130 (30/100) / 53%	Full Face, Audio	HCRF	F1=0.644	0.58 [*3]
	Nasir et al. (2016) [151]	35 (7/28) / 45.7%	Facial Landmarks	SVM	F1=0.63	0.55 [*3]
	Pampouchidou et al. (2016) [150]	-//-	Facial Landmarks, Audio	Random Forest	F1=0.62	0.53
	Valstar et al. (2016) [96]	-//-	Facial Landmarks, AUs, Audio	SVM	F1=0.5	0.45[*3]
	Valstar et al. (2016) [96]	-//-	Facial Landmarks, AUs	SVM	F1=0.5	0.45 [*3]
	Pampouchidou et al. (2016) [150]	-//-	Facial Landmarks	Random Forest	F1=0.5	0.39
	Scherer et al. (2013) [98]	39 (14/25) / 62%	Full Face	SVM	64.1%	0.20 [*3]
AVEC	Senoussaoui et al. (2014) [129]	50 (25/25)	Full Face	SVM	82%	0.64
	Alghowinem et al. (2015) [106]	32 (16/16)/ 28.1%	Eye, Full Face	SVM	mean recall =68.8%	0.38
	Pampouchidou et al. (2016) [144]	200 (34/166) [*1] / 33%	Full Face	Nearest Neighbour	74.5%	0.13

[*1]: *Reported sample size corresponds to number of sessions.*
[2]: *Three classes were considered, corresponding to the annotation in the table as follows: ([Moderate to Severe]/[Mild]/[Remission]).*
[3]: *The confusion matrix was computed based on performance metrics provided by the authors of the original report.*

Where $precision$ is given by

$$precision = \frac{TP}{TP + FP}, \tag{5}$$

and $recall$ by

$$recall = \frac{TP}{TP + FN}. \tag{6}$$

The F1 score, also reported in several studies, is given by

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{7}$$

The aforementioned performance metrics, however, fail to take chance agreement into consideration, which varies across different studies. To address this limitation and to

TABLE 8
Approaches for Categorical Assessment of Depression Employing Datasets, or Combination of Datasets, Which Have Not Been Reported Elsewhere

Data	Paper	Population (Study/ Control) / Male rate	Features	Classification Algorithm	Reported Accuracy (or precision)	Kappa
Pittsburgh + AVEC'14	Alghowinem et al. (2015) [106]	70 (35/35) / 32.9%	Eyes, Full Face	SVM	mean recall = 85.7%	0.57
Rochester	Zhou et al. (2015) [75]	10 (5/5) /-	Full Face, Eyes	Logistic regression	precision = 0.82	0.57
ORI	Maddage et al. (2009) [138]	8 (4/4) / 50%	Full Face	GMM	75.6%	0.45*
ORYGEN	Ooi et al. (2011) [117], [125]	30 (15/15)/ 51%	Full Face	Nearest Neighbour	51%	0.03*
CHI-MEI	Yang et al. (2016) [95]	26 (13/13) / -	AUs, Audio	CHMM	65.38%	0.26
	Yang et al. (2016) [95]	26 (13/13) / -	AUs	HMM	53.85%	0.08

** Confusion matrix was computed, and not provided by the authors of the original research report.*

permit direct comparisons between classification approaches, Cohen's Kappa statistic [169], a chance and skew robust metric, was computed whenever possible. It is based on the confusion matrix (3), and given by

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \tag{8}$$

where p_0 is the proportion of accurately predicted decisions given by the accuracy formula as defined in (4), and p_e the proportion of expected chance agreement, given by

$$p_e = \frac{M_a + M_b}{TP + FN + FP + TN}, \tag{9}$$

where M_a and M_b are defined as follows:

$$M_a = \frac{(TP + FN) * (TP + FP)}{TP + FN + FP + TN} \tag{10}$$

$$M_b = \frac{(TN + FP) * (TN + FN)}{TP + FN + FP + TN}. \tag{11}$$

Whenever confusion matrices for user/gender independent depression assessment (based on one or more visual cues) were not included in the original publication, they were requested from the study authors. For the cases that the requested information was not provided, and if at least two performance metrics where reported in the original publication, along with the total number of subjects per class (depressed, non-depressed as defined in (12) and (13)), a 4×4 linear system of equations was solved in order to derive the confusion matrix

$$\#depressed = TP + FN \tag{12}$$

$$\#not\text{-}depressed = TN + FP. \tag{13}$$

In the case of [98] a quadratic system was solved, since the reported metrics were averaged for the two classes (depressed/not-depressed). Finally, the computed confusion matrices were cross-checked to reproduce the originally reported performance metrics. If the estimated confusion matrices for a given study could not be verified by the reported performance metrics, the relevant study was not considered any further. It should be noted that Dibeklioğlu et al. [164] is the only reviewed publication which originally included Kappa statistics in their published report.

Table 7 groups the reviewed studies according to the dataset used, the corresponding sample size and male rate. The table also lists the classification algorithm(s) used in each study, the features employed in the decision process, and the corresponding performance metric. Grouping of studies was inspired by [134]. Studies are ranked by decreasing κ value within each dataset-specific group. Table 8 presents similar information on studies using datasets or dataset combinations reported in single studies, precluding direct across-study comparisons. Finally, Fig. 8 displays a forest plot of κ values with the corresponding 95 percent confidence intervals given by

$$95\%CI = \kappa \pm 1.96 * \sigma_\kappa, \tag{14}$$

with σ_κ defined as

$$\sigma_\kappa = \sqrt{\frac{p_0(1 - p_0)}{N(1 - p_e)^2}}, \tag{15}$$

where N is the total number of samples. Below we attempt an assessment of the evidence provided by this meta-analysis.

5.1.1 Pittsburgh

The first report involving the Pittsburgh dataset is that of Cohn et al. [82], who presented results of three experiments on depression detection based on: a) manual annotation of AUs using AU14 (lip corner tightening), b) automatic detection with AAMs, and c) vocal prosody. They reported 88 percent accuracy in experiment (a), and 79 percent in both (b) and (c), stressing the need to adapt AAMs to individual patients [88]. The authors conclude that non-verbal, vocal cues constitute a valid indicator of depression severity [92].

Girard et al. [90] extended this work by investigating the correlation of changes in patient clinical status with corresponding changes in facial expression and head motion patterns over four sessions at six-week intervals. They concluded that as depression severity was reduced, smiling became more frequent and expressions of contempt and embarrassment became less frequent [91].

The cross-cultural study of Alghowinem et al. [106] was also tested on the Pittsburgh dataset among others, reporting an average recall of 94.7 percent. Dibeklioğlu et al. [136] tested several feature settings on the Pittsburgh dataset. More recently Dibeklioğlu et al. [164] presented a deep learning approach for detecting three levels of depression. Finally, Joshi et al. [113], and Joshi [110], reported 97.2

TABLE 9
Summary of Methods and Results of Studies Employing Continuous Depression Assessment Based on the Dataset Provided by AVEC'13, AVEC'14 and AVEC'16

Paper	Regression	Fusion	Development		Test	
			Visual	Multimodal	Visual	Multimodal
AVEC 2013 - BDI prediction - Complete recordings						
Zhu et al. (2017) [149]	DCNN	–	–	–	7.58 / 9.82	–
Kaya et al. (2014) [142]	MPGI	Decision	–	–	8.254 / 10.315	7.693 / 9.611 ↑
Cummins et al. (2013) [122]	SVR	Feature	– / 12.08	– / 10.44 ↑	– / 10.45	– / 10.62 ↓
Meng et al. (2013) [121]	PLS	Decision	7.09 / 8.82	6.94 / 8.54 ↑	9.14 / 11.19	8.72 / 10.96 ↑
Valstar et al. (2013) [51]	SVR	–	8.74 / 10.72	–	10.88 / 13.61	–
AVEC 2014 - BDI prediction - Northwind / Freeform tasks						
Zhu et al. (2017) [149]	DCNN	–	–	–	7.47 / 9.55	–
Kaya & Salah (2014) [145]	CCA	Decision	–	–	7.86 / 9.72	7.68 / 9.44 ↑
Jain et al. (2014) [139]	SVR	Feature	6.969 / 8.167	6.964 / 8.162 ↑	8.399 / 10.249	-
Jan et al. (2014) [124]	PLS+LR	Decision	7.36 / 9.49	7.34 / 9.09 ↑	8.44 / 10.50	8.30 / 10.26 ↑
Kächele et al. (2014) [127]	SVR	Decision	7.03 / 8.82	8.30 / 9.94 ↓	8.97 / 10.82	9.09 / 11.19 ↓
Valstar et al. (2014) [53]	SVR	Decision	7.577 / 9.314	6.680 / 8.341 ↑	8.857 / 10.859	7.893 / 9.891 ↑
Senoussaoui et al. (2014) [129]	GLM + RVM + SVR	Decision	6.95 / 8.52	6.57/7.91 ↑	-	8.33 / 10.43
Smailis et al. (2016) [137]	SVR	Feature	- / 9.07	- / 9.16 ↓	–	–
He et al. (2015) [148]	SVR	Decision	7.99 / 9.63	6.17 / 7.67 ↑	–	–
Sidorov & Minker (2014) [126]	SVR	Feature	14.843 / 17.667	7.245 / 8.964 ↑	-	8.327 / 10.826
Williamson et al. (2014) [93]	GMM+ELM	Decision	–	– –	–	– / 8.12
AVEC 2016 - PHQ-8 prediction - DAIC-WOZ						
Valstar et al. (2016) [96]	RFR	Decision	5.88 / 7.13	5.52 / 6.62 ↑	6.12 / 6.97	5.66 / 7.05 ↑↓
Yang et al. (2016) [146] (females)	LLR	Decision	3.26 / 3.97	2.63 / 3.77 ↑	–	–
Yang et al. (2016) [146] (males)	SVR	Decision	3.19 / 4.29	2.94 / 3.67 ↑	–	–
Williamson et al. (2016) [167]	GSM	Decision	5.33 / 6.45	4.18 / 5.31 ↑	–	–

Performance metrics are given in the form of MAE / RMSE and approaches have been ranked according to reported performance primarily on Test-Visual-RMSE, and second on Development-Visual-RMSE.

percent accuracy for assessing depression severity based on the Pittsburgh data.

5.1.2 BlackDog

McIntyre et al. [83] were the first to report on the BlackDog dataset. Their results support the long-standing clinical observation that depressed individuals demonstrate lesser facial activity and a smaller repertoire of facial expressions. They also reported identifying two clusters of patients, those who showed psychomotor agitation and those who showed motor slowing [87], [89].

In a subsequent study, Joshi et al. [112] focused on algorithm development and improvement and compared the performance of different neural network classification

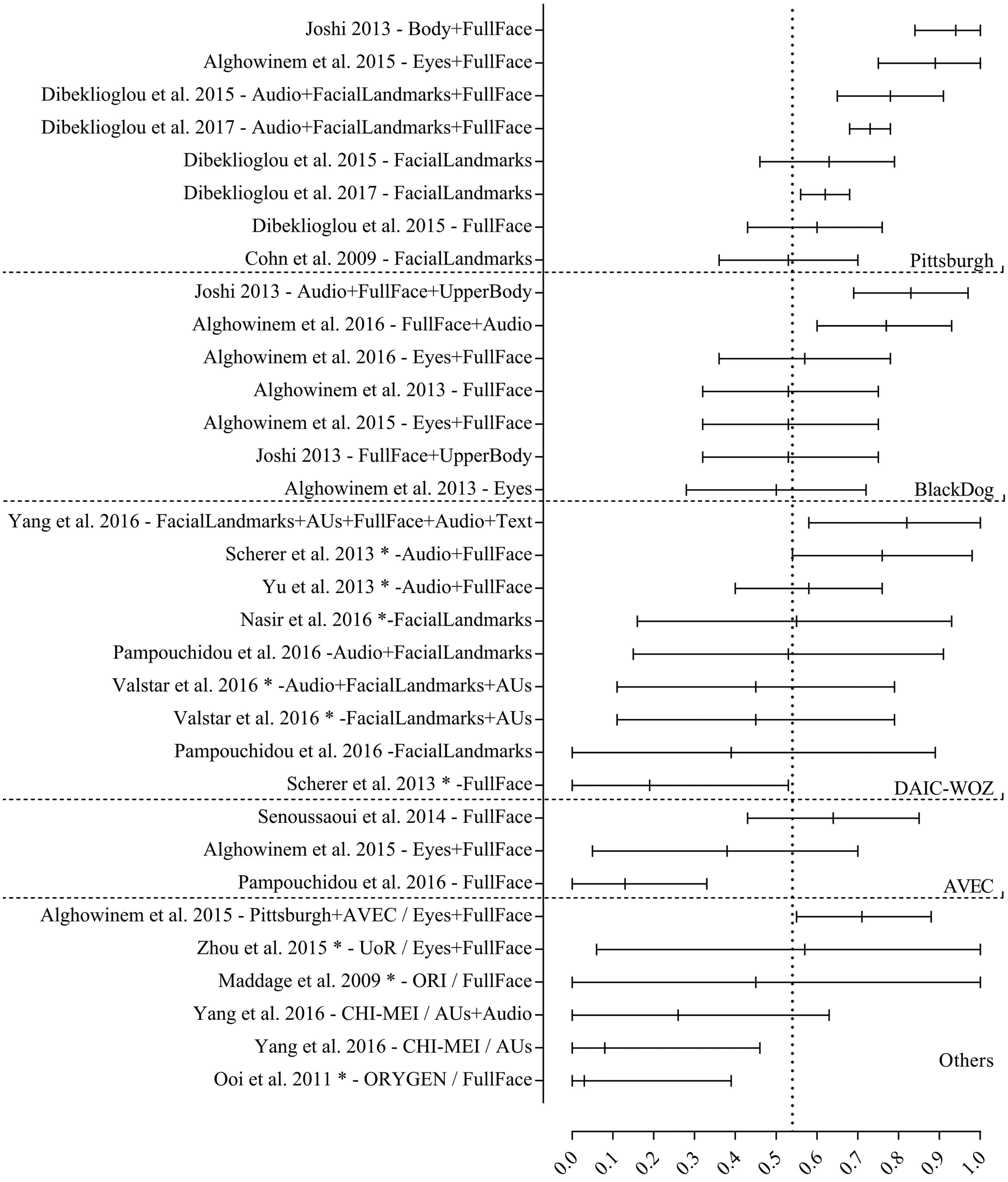

Fig. 8. Forest plot of study-specific κ values, grouped by dataset. The vertical dashed line corresponds to the average $\kappa = 0.54$.

algorithms achieving depression detection accuracy as high as 88.3 percent. Higher performance, up to 91.7 percent, was achieved when additional modalities were included (speech, independent and relative movement of body parts) in Joshi et al. [111] and Joshi et al. [123].

Alghowinem et al. studied depression detection based on the analysis of either eye movements in [105], or head pose and head movements in [108] by extracting 126 and 100 features, respectively. The maximum reported recall rate was 80 percent for the eye-based approach, and 82.6 percent for the head-based approach among women (corresponding rates were lower for men: 77 and 75.6 percent, respectively). In the same cross-cultural study mentioned in Section 5.1.1, Alghowinem et al. [106] combined the two approaches (eye-based and head-based) achieving an average recall of 85 percent for a variable set of features (made specific for the BlackDog), and 76.7 percent for a fixed set of features along the different datasets employed in their cross-cultural study. Recently, Alghowinem et al. [116] presented an improved performance of recall 88.3 percent, by extending their approach in terms of feature extraction, as well as machine learning methods.

5.1.3 DAIC-WOZ

The creation of the DAIC-WOZ dataset was based on the SimSensei kiosk tool, developed at the Institute of Creative Technologies, University of Southern California (ICT-USC) [100], [102], [170], [171], [172]. SimSensei is a virtual human dialogue system created to conduct personal interviews while recording and analyzing the facial image and speech of the interviewee. Several approaches have been tested on this dataset, mainly from the relevant research group, but also in AVEC'16, in terms of the depression sub-challenge.

Scherer et al. [101] examined the value of the audiovisual approach, achieving 89.74 percent accuracy compared to 51.28 percent of single-modality acoustic features, and 64.10 percent for the single-modality visual features. In subsequent publications, Scherer et al. [98], [103], examined linear associations between frequency of specific features and disorder type, showing significant differences between Post Traumatic Stress Disorder (PTSD), anxiety, depression, and distress on gaze behavior, smiling, self-touching and fidgeting.

Stratou et al. [84] corroborated Alghowinem's findings [105], [108] on gender differences in classification accuracy, reporting F1 scores of 0.858 for women and 0.808 for men in detecting presence of depression and PTSD. Yu et al. [99] surmised that extracting and integrating features over the entire interview is not as accurate as extracting features from separate question-answer instances ("adjacency-pairs"). Results for the adjacency-pairs approach revealed improved performance as indicated by an F1 score of 0.603 compared to a score of 0.523 for the data integrated over the entire session. The MaxEnt algorithm was used in both approaches.

Ghosh et al. [85] integrated features derived from facial images and speech with affect (sadness, anxiety, anger) displayed in a chat-text and achieved improved classification accuracy of 66.40 percent as compared to 63.02 percent for unimodal text, 58.63 percent for speech, 58.00 percent for facial image, and 60.50 percent for the combination of facial image and speech.

Finally, DAIC-WOZ served as the benchmark dataset in terms of AVEC'16. Raw data were provided for audio recordings and transcripts, while for videos only features extracted with OpenFace were provided. Despite this limitation several interesting approaches were presented, with Yang et al. [146] winning the depression sub-challenge by combining visual with audio and text features.

5.1.4 AVEC

The AVEC dataset, although intended for continuous assessment of depression, was also used for categorical assessment using case groupings based on BDI scores. For instance, Senoussaoui et al. [129] achieved classification accuracy of 82 percent for categorical assessment of depression by using a cutoff score of 13/14 points on BDI, using the data organization provided by AVEC (training / development subsets).

Alghowinem et al. [106], in their cross-cultural study tested their algorithm on a subset of the AVEC data, in an effort to match the three datasets employed (Pittsburgh, BlackDog, and AVEC), both in terms of number of recordings as well as total duration. They reported an average recall of 68.8 percent for the fixed set of features across the three datasets.

Pampouchidou et al. [144] considered standard BDI cutoffs: 0-9 (minimal depression), 10-18 (mild depression), 19-29 (moderate depression), and 30-63 points (severe depression). They reported 74.5 percent accuracy in discriminating cases of {minimal} versus {mild, moderate, severe}, and {minimal, mild, moderate} versus {severe} depressive symptomatology. The classification accuracy between {minimal, mild} versus {moderate, severe} was 63.5 percent.

5.1.5 Other Datasets

Datasets or data set combinations used in a single study are summarized in Table 8. Alghowinem et al. [106] attempted to merge several datasets (e.g., Pittsburgh and AVEC). This resulted in an improvement of classification performance as compared to relying solely on the AVEC dataset, reporting an average recall of 85.7 percent.

In one of the earliest studies, the corpus constructed by the Oregon Research Institute (ORI), was used to test the approach of Maddage et al. [138] for video-based depression detection in adolescents. The corpus comprised recordings from eight adolescents (four in each class). They implemented both gender-dependent and independent classification achieving 75.6 percent recognition rate of adolescents with depression in the gender-independent case.

The ORYGEN dataset was employed for a more challenging endeavor undertaken by Ooi et al. [117]. Facial expression analysis was utilized in order to predict whether initially non-depressed adolescents would develop depression at the end of a two-year follow-up period. They obtained baseline (Time 1) and two-year follow-up (Time 2) data from 191 non-depressed adolescents. At Time 2, 15 participants had developed depression. Given the still-disputed capacity of available computer based approaches to detect the current presence of depression, long-term prediction of depression onset was a rather optimistic goal. As expected, results revealed relatively poor prediction accuracy (50 percent for person-independent, and 61 percent for person-dependent) casting doubt on the sensitivity of facial features as prodromal signs of clinical depression [125].

Zhou et al. [75] at the University of Rochester departed from the traditional laboratory settings and obtained data in realistic conditions. Participants interacted with each other through social networking, while sitting in an area customized to resemble a living room. Pupil radius, head movement rate, eye blinking rate, text-derived affect, and keystroke rate are some of the cues they considered for detecting the presence of depressive symptomatology. They reported 0.817 precision and 0.739 recall for classifying patients versus healthy volunteers reporting high levels of negative mood. Although promising, the generalizability of these results is questionable given the small number of patients with depression (n = 5).

Finally, recent results from the CHI-MEI dataset [95], attempting to distinguish unipolar depression (MDD) from bipolar disorder, reached 65.38 percent classification accuracy when combining AUs and audio features.

5.2 Continuous Depression Assessment

The majority of the approaches reviewed here are based on AVEC datasets as participations to the actual challenge, or as independently published studies. The depression challenge of AVEC 2013 and AVEC 2014 required prediction of individual BDI scores based on corresponding video recordings (both visual and speech data are available). Video recordings were divided into three subsets (training, development and testing), with labels being released only for the first two. AVEC 2013 included the complete recordings of the participants executing 12 tasks, while for AVEC 2014 only two tasks were kept: the Northwind (reading a novel excerpt) and Freeform (answering to an open question). AVEC 2016, although focused on categorical assessment, encouraged participants to address prediction of self-reported scores on the PHQ-8 scale, employed by DAIC-WOZ.

In all cases, performance metrics were the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) given by Equations (16) and (17), where n is the number of samples, p the predicted value, and a the actual value

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - a_i)^2} \tag{16}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - a_i|. \tag{17}$$

The MIT-Lincoln team won both 2013 and 2014 challenges [93], [173]. In their first participation they achieved subject-independent RMSE/MAE of 8.5/6.53 on the test set, 8.68/7.12 on the development, and 7.42/5.75 subject-dependent adaptation on the development set, relying solely on vocal features. By incorporating AU-based features, the team achieved slightly better results in the subsequent challenge on a different subset (8.12/6.31 on the test set; results for the development set were not reported). In a recent article, Zhu et al. [149], presented a deep learning approach tested on the datasets provided by AVEC'13 and AVEC'14. Their method was single-visual and deep learning based, outperforming other reported single-visual approaches, and comparing favorably with the overall best performing (multimodal) results (cf. Table 9), with RMSE/MAE of 9.55/7.47 for AVEC'14, and 9.82/7.58 for AVEC'13 respectively. In terms of AVEC 2016, Williamson et al. [167] and Yang et al. [146] addressed prediction of PHQ-8, with the latter predicting scores separately for females and males.

Table 9 compares the performance of each algorithm when tested on visual versus multimodal signals. The regression algorithm, along with the fusion strategy and the respective scores, for both development and test subsets, are also presented where available. Upward arrows indicate better performance of the multimodal versus the visual approach alone and the opposite is indicated by downward arrows.

6 Discussion

Concluding comments are organized according to the datasets and algorithms used in each of the reviewed studies.

6.1 Algorithm and Performance Related Issues

Given the temporal variability of depressive manifestations, the majority of facial signs utilized in the reviewed studies are dynamic, while the occurrence of static signs is typically considered over time (e.g., smile frequency). Therefore, video recordings, as opposed to static images, are required. Further, a trend toward utilizing high-level features has been observed; this trend was also promoted by AVEC'16, in which features were provided after preprocessing, that enabled high-level feature extraction. Furthermore, the majority of top performing methods employ multiple sets of features, across or even within a given modality, e.g., visual cues from face and body.

With respect to the continuous assessment approaches, multimodal methods (audio/visual) clearly outperform those relying on solely on visual cues. Thus, as indicated by the arrows in Table 9, the inclusion of audio cues resulted in performance improvement in 17 out of 22 methods reporting on both single-visual and multimodal performance. Furthermore, decision fusion appears to be the most prominent. Potential benefits from multimodal approaches have also been discussed in [64]. Finally, gender-based classification has been supported by many of the reported approaches to perform better [86], [116], [138], [146], [150].

Deep learning based approaches have been found in only two recent articles. In the first, Dibeklioğlu et al. [164] attempted a multimodal approach, employing Stacked Denoising Autoencoders for detecting three levels of depression, and reported comparable results to two-level approaches. Also, Zhu et al. [149] addressed the AVEC'13 and AVEC'14 depression sub-challenges using Deep Convolutional Neural Networks, achieving the highest performance among the single-visual approaches, although underperforming the multimodal ones. Deep learning seems to be promising, demanding further exploration. It is worth noting that multimodal approaches are again performing best.

Treating depression as a continuous variable (i.e., reflecting depression severity) is gaining ground over categorical decision systems (e.g., as reflected in AVEC'13 and AVEC'14). This is due to the fact that, despite the apparent simplicity of categorical assessment, it does not represent neither properly nor reliably the complex nature of mood disorders. Efforts to develop methods capable of differentiating mood disorder types, and also depression from other psychiatric disorders, such as various anxiety conditions, are limited. However, there have been some notable initial attempts in this direction by ICT-USC focusing on distinguishing PTSD from depression [84], [86], and by CHI-MEI attempting to distinguish cases of unipolar depression versus bipolar disorder [95]. It is worth pointing out that, regarding long-term prediction of depression onset, available results reveal relatively poor prediction accuracy calling into question the significance of facial manifestations as prodromal signs of depression.

Finally, although reported detection accuracy rates can be very high, a fact that clearly demonstrates the clinical potential of the field, sample sizes are often too small to enable the generalizability of these results. In order for a system to be fully evaluated and acknowledged as an assessment tool, it must be tested on considerably larger sample sizes, featuring a wider variety of demographic characteristics, clinical diagnosis methods, and ethnic-cultural backgrounds.

6.2 Data Related Issues

An initial observation based on the meta-analysis performed in Section 5.1, summarized by Tables 7 and 8, and Fig. 8, permits a rough comparison of methods, through Cohen's κ (a chance-robust metric). Approaches tested on the Pittsburgh dataset achieved higher (in their majority above average) performance. Similar results were reported by Alghowinem et al. [106], who tested the exact same algorithm on different datasets (cf. Tables 7 and 8, and Fig. 8). This finding highlights the importance of data quality (sample size, noise, resolution, environment, etc.) for the development and testing of a given computational pipeline. Further, all participants in the Pittsburgh dataset met DSM criteria for clinical depression, ensuring diagnostic uniformity and exclusion of potentially important comorbid conditions [164]. Finally, the data were obtained during a clinical interview, thus enhancing the interpersonal context [164].

Regarding the categorical approaches tested on the AVEC and DAIC-WOZ datasets, there are some additional potential reasons which could have affected the performance, as revealed through comparison of κ values cross studies. First, inclusion of participants from different ethnicities than other datasets may have biased the result of such experimental manipulations as mood induction. Further, the AVEC and DAIC-WOZ datasets were collected in a way that limited the audience effect (human-computer interaction setup), eliminating cues that could only occur in a social context as, for instance, indicated by Dibeklioğlu et al. [164]. Also, the Pittsburgh and BlackDog datasets were both based on clinical diagnosis, as opposed to AVEC and DAIC-WOZ, which were based on self-reports; a factor expected to affect annotation accuracy [116].

Most importantly, the problem addressed with the AVEC and DAIC-WOZ datasets, which relied on self-reported symptoms for data annotation, is far more complex in comparison to the other datasets. While others focused on categorical assessment (healthy versus depressed, high versus low depression severity) based on clinical diagnosis, AVEC focused on the prediction of individual BDI scores. Even during AVEC'16, where DAIC-WOZ was used for categorical assessment, binary labels (depressed/not-depressed) were again based on participant symptom self-report (PHQ-8 scores). As explained in the Clinical Background section, Self-RI scores depend on a variety of biasing factors (subjective, social, etc.). Thus, although depression can be accurately portrayed by facial expressions, the ground-truth may not accurately measure depression severity in these studies. This could also justify the fact that reported approaches for categorical assessment outnumber the ones for continuous assessment, as the continuous prediction is more challenging.

So what constitutes an optimal data set? It appears that it should be comprised of a number of patients diagnosed by experienced psychiatrists, using largely uniform diagnostic criteria. Comorbid diagnoses, should be carefully recorded and later used to evaluate potential misclassifications, since significant correlations have been observed between PTSD, anxiety and depression [103]. Consequently, the development of an algorithmic tool for depression assessment, apart from the contribution of engineers and computer scientists, would most definitely require direct supervision by clinicians. Furthermore, as reported in the clinical background of Section 1.1, a one-off assessment may not be sufficient, as development of rapport with the participant is necessary. For this to be achieved several sessions over a fixed interval (e.g., 7 weeks as in [82], [90], [91]) are advisable. Existence of baseline data would also be useful, but unfortunately this is not possible in most cases. However multiple sessions can benefit the remission assessment, allowing long-term monitoring of the recovery process.

With respect to video acquisition parameters, average image resolution and frame raterate are typically reported. However, it is not clear whether image acquisition with higher-level specifications could improve assessment accuracy. A quantitative comparison of different resolutions and frame rates employing the same algorithm(s) may be required to conclusively address this question. The size of the sample is also an important factor to be considered, as size of the majority of reported datasets is at best moderate to allow generalizations [116].

In the field of automatic facial expression recognition (AFER) approaches are moving toward real-world conditions [62], as exemplified by the Emotion Recognition in-the-Wild (EmotiW) challenge series [174], [175], [176]. The manner in which the AVEC dataset was constructed also supports this idea, as the recordings took place in independent setups and on personal computers. This choice, however, impacted performance, as shown in Table 7 and Fig. 8: approaches for categorical assessment of depression based on AVEC demonstrated lower than average performance, when compared to approaches based on other datasets. Additionally, although current in-the-wild approaches may be considered as promising, they are not yet sufficiently reliable even for AFER as supported in [61], [62]. Therefore, at present, such approaches do not appear to meet minimum requirements for a clinical decision support system. On the other hand, the strict requirement for standardization of data collection [64] may impose potentially serious limitations, such as questionable originality and genuineness of the data and lack of variance in contextual information. Although standardized medical equipment typically operate under highly controlled conditions, collection of data indicative of depressive symptomatology is highly susceptible to the dynamic nature of behavioral and underlying psychological processes of the person being evaluated.

The most important issue concerning the data is availability. As mentioned before, obscuring participant identity is practically impossible in raw video data compared to other modalities (e.g., speech or physiological signals). Consequently, open access is strictly prohibited, while licensing to a third party is seriously restricted. A possible solution to this problem would be for the academic community to promote cooperation between institutions, as suggested by Cummins et al. [64], so that the data could be shared under regulated conditions.

7 Conclusions & Future Work

Research on automatic depression assessment has come a long way from Cohn et al. [82] and McIntyre et al. [83], with several novel approaches both in terms of methodology and performance. The present comprehensive review of the state-of-the-art provides a number of insights, while identifying many questions open to further investigation. Depression diagnosis itself is an active and controversial topic in clinical psychology and psychiatry. Given the aforementioned outstanding issues, the development of automated, objective assessment methods may be valuable for both research and clinical practice.

In general, results are consistent with the social withdrawal [86], [90], emotion-context insensitivity [90], and reduced reactivity [164] hypotheses of depression. Additionally, the importance of dynamic features, as well as multimodal approaches, was also highlighted through the quantitative analysis reported in this review. Continuous approaches appear to be more in accordance with clinical experience. Sharing of related data needs to be promoted, and the data collection procedures need to be standardized on several domains, in order to enhance reliability and comparability of results. In terms of related algorithms, although a multitude of approaches is reported in the literature, automatic depression assessment is still a long way from being well established, with significant room for improvement on current methods. Given the dynamic of deep learning, and its considerably high performance in many related fields, if large enough datasets are provided, an impact in automatic depression assessment is also expected in the near future.

Several clinical research questions remain to be addressed systematically, such as the capacity to distinguish between different depression subtypes, and MDD from other mood disorders [86]. Individual variability due to comorbid personality disorders or characteristics, as well as the influence of ethnicity and culture requires further exploration [116]. Interestingly, although physiological activity measured through EMG, BVP, skin conductance, and respiration can be informative regarding ongoing emotional responses [177], such information has not been integrated in the reviewed multimodal studies, with the exception of Zhou et al. [75], who recorded heart rate via a non-contact, facial video-based system. Accordingly, body manifestations, as well as pupil related features, have not been adequately employed for automatic assessment, although they have been proven statistically significant. Finally, investigation of facial signs in the context of dyadic interaction is a forthcoming domain for exploration [103], still only one approach considered it [99], while context has not been considered in any study.

In conclusion, the exhaustive review of available evidence highlights the considerable potential in the application of video-based methods for the assessment and monitoring of the course of depression. It was further made apparent that visual cues need to be supplemented by information from other modalities to achieve clinically useful results.

Acknowledgments

Anastasia Pampouchidou was funded by the Greek State Scholarship Foundation, under the scholarship programme instituted in memory of Maria Zaousi. This work was also partially supported by a grant entitled SEMEOTICONS (SEMEiotic Oriented Technology for Individual's CardiOmetabolic risk self-assessmeNt and Self-monitoring) within the framework of the FP7 Specific Targeted Research Project partially funded by the European Commission under Grant Agreement 611516. The authors would further like to express their gratitude to Prof. Wu and to Drs. Alghowinem, Joshi, Dibeklioğlu, and Senoussaoui for providing unpublished confusion matrices of their approaches. Finally, the authors express their gratitude to the anonymous reviewers of earlier versions of this work for their constructive comments and suggestions.

References

[1] World Health Organization, May 2017. [Online]. Available: http://www.who.int/mental_health/management/depression/en/

[2] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. Arlington, VA, USA: American Psychiatric Publishing, 2013.

[3] R. C. Kessler, "The costs of depression," *Psychiatric Clinics North America*, vol. 35, no. 1, pp. 1–14, 2012.

[4] J. C. Wakefield and S. Demazeux, "Introduction: Depression, one and many," in *Sadness or Depression? International Perspectives on the Depression Epidemic and Its Meaning*. Dordrecht, The Netherlands: Springer, 2016, pp. 1–15.

[5] Survey of Health, Ageing and Retirement in Europe, Feb. 2015. [Online]. Available: http://www.share-project.org

[6] P. van de Ven, M. R. Henriques, M. Hoogendoorn, M. Klein, E. McGovern, J. Nelson, H. Silva, and E. Tousset, "A mobile system for treatment of depression.," *Comput. Paradigms for Mental Health*, vol. 47, 2012.

[7] K. N. Fountoulakis, et al., "Relationship of suicide rates to economic variables in Europe: 2000–2011," *Brit. J. Psychiatry*, vol. 205, no. 6, pp. 486–496, 2014.

[8] P. Sobocki, B. Jönsson, J. Angst, and C. Rehnberg, "Cost of depression in Europe," *J. Mental Health Policy Economics*, vol. 9, no. 2, pp. 87–98, 2006.

[9] J. Olesen, et al., "The economic cost of brain disorders in Europe," *Eur. J. Neurology*, vol. 19, no. 1, pp. 155–162, 2012.

[10] R. C. Kessler and E. J. Bromet, "The epidemiology of depression across cultures," *Annu. Rev. Public Health*, vol. 34, pp. 119–138, 2013.

[11] D. Goldberg, "The current status of the diagnosis of depression," in *Sadness or Depression? International Perspectives on the Depression Epidemic and Its Meaning*. Dordrecht, The Netherlands: Springer, 2016, pp. 17–27.

[12] J. A. Blalock and T. J. E. Joiner, "Interaction of cognitive avoidance coping and stress in predicting depression/anxiety," *Cogn. Therapy Res.*, vol. 24, no. 1, pp. 47–65, 2000.

[13] M. B. First and M. Gibbon, *The Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) and the Structured Clinical Interview for DSM-IV Axis II Disorders (SCID-II)*. M. J. Hilsenroth and D. L. Segal (Eds.), Comprehensive handbook of psychological assessment, vol. 2. Personality assessment. Hoboken, NJ: John Wiley, 2004, pp. 134–143.

[14] M. Hamilton, "A rating scale for depression," *J. Neurology Neurosurgery Psychiatry*, vol. 23, no. 1, 1960, Art. no. 56.

[15] R. M. Bagby, A. G. Ryder, D. R. Schuller, and M. B. Marshall, "The hamilton depression rating scale: Has the gold standard become a lead weight?" *Amer. J. Psychiatry*, vol. 161, no. 12, pp. 2163–2177, 2004.

[16] M. Chmielewski, L. A. Clark, R. M. Bagby, and D. Watson, "Method matters: Understanding diagnostic reliability in DSM-IV and DSM-5," *J. Abnormal Psychology*, vol. 124, no. 3, 2015, Art. no. 764.

[17] H. C. Kraemer, D. J. Kupfer, D. E. Clarke, W. E. Narrow, and D. A. Regier, "DSM-5: How reliable is reliable enough?" *Amer. J. Psychiatry*, vol. 169, no. 1, pp. 13–15, 2012.

[18] R. Thomas-MacLean, J. Stoppard, B. B. Miedema, and S. Tatemichi, "Diagnosing depression: There is no blood test," *Can. Family Physician*, vol. 51, no. 8, pp. 1102–1103, 2005.

[19] S. Kapur, A. G. Phillips, and T. R. Insel, "Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it?" *Mol. Psychiatry*, vol. 17, no. 12, pp. 1174–1179, 2012.

[20] I. Schumann, A. Schneider, C. Kantert, B. Lwe, and K. Linde, "Physicians attitudes, diagnostic process and barriers regarding depression diagnosis in primary care: A systematic review of qualitative studies," *Family Practice*, vol. 29, no. 3, pp. 255–263, 2011.

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

[22] K. E. Markon, M. Chmielewski, and C. J. Miller, "The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review," *Psychological Bulletin*, vol. 137, no. 5, 2011, Art. no. 856.

[23] M. Maj, "The continuum of depressive states in the population and the differential diagnosis between "Normal"' sadness and clinical depression," in *Sadness or Depression? International Perspectives on the Depression Epidemic and Its Meaning*. Dordrecht, The Netherlands: Springer, 2016, pp. 29–38.

[24] L. S. Williams, et al., "Performance of the PHQ-9 as a screening tool for depression after stroke," *Stroke*, vol. 36, no. 3, pp. 635–638, 2005.

[25] P. Pichot, "Self-report inventories in the study of depression," in *New Results in Depression Research*. Berlin, Germany: Springer, 1986, pp. 53–58.

[26] C. Cusin, H. Yang, A. Yeung, and M. Fava, "Rating scales for depression," in *Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health*. Berlin, Germany: Springer, 2009, pp. 7–35.

[27] Y. S. Ben-Porath, "Assessing personality and psychopathology with self-report inventories," in *Handbook of Psychology*. Hoboken, NJ, USA: Wiley, 2003, pp. 553–577.

[28] S. Gilbody, T. Sheldon, and A. House, "Screening and case-finding instruments for depression: A meta-analysis," *Can. Med. Assoc. J.*, vol. 178, no. 8, pp. 997–1003, 2008. [Online]. Available: http://www.cmaj.ca/content/178/8/997.abstract

[29] Y. Ren, H. Yang, C. Browning, S. Thomas, and M. Liu, "Performance of screening tools in detecting major depressive disorder among patients with coronary heart disease: A systematic review," *Med. Sci. Monitor*, vol. 21, pp. 646–653, 2015. [Online]. Available: www.medscimonit.com/abstract/index/idArt/892537

[30] E. Stockings, et al., "Symptom screening scales for detecting major depressive disorder in children and adolescents: A systematic review and meta-analysis of reliability, validity and diagnostic utility," *J. Affective Disorders*, vol. 174, pp. 447–463, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016503271400785X

[31] A. J. Mitchell and J. C. Coyne, *Screening for Depression in Clinical Practice: An Evidence-Based Guide*. London, U.K.: Oxford Univ. Press, 2009.

[32] C. Hollis, R. Morriss, J. Martin, S. Amani, R. Cotton, M. Denis, and S. Lewis, "Technological innovations in mental healthcare: Harnessing the digital revolution," *Brit. J. Psychiatry*, vol. 206, no. 4, pp. 263–265, 2015.

[33] J. M. Girard and J. F. Cohn, "A primer on observational measurement," *Assessment*, vol. 23, pp. 404–413, 2016.

[34] J. M. Girard and J. F. Cohn, "Automated audiovisual depression analysis," *Current Opinion Psychology*, vol. 4, pp. 75–79, 2015.

[35] ACM. [Online]. Available: http://dl.acm.org, Accessed in May 2015.

[36] IEEE. [Online]. Available: http://ieeexplore.ieee.org, Accessed in May 2015.

[37] Elsevier. [Online]. Available: http://www.sciencedirect.com, Accessed in May 2015.

[38] Springer. [Online]. Available: http://link.springer.com, Accessed in May 2015.

[39] Wiley. [Online]. Available: http://onlinelibrary.wiley.com, Accessed in May 2015.

[40] NASA. [Online]. Available: http://lsda.jsc.nasa.gov, Accessed in May 2015.

[41] Oxford. [Online]. Available: http://www.oxfordjournals.org, Accessed in May 2015.

[42] PubMed. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed, Accessed in May 2015.

[43] Scopus. [Online]. Available: http://www.scopus.com, Accessed in May 2015.

[44] GoogleScholar. [Online]. Available: http://scholar.google.gr, Accessed in May 2015.

[45] MedPilot. [Online]. Available: http://www.medpilot.de, Accessed in May 2015.

[46] Mayo Clinic, Feb. 2015. [Online]. Available: http://www.mayoclinic.org/diseases-conditions/depression/basics/tests-diagnosis/con-20032977

[47] National Comorbidity Survey - Harvard Medical School, Sep. 2016. [Online]. Available: http://www.hcp.med.harvard.edu/ncs/

[48] World Health Organization - Regional Office for Europe, Feb. 2015. [Online]. Available: http://www.euro.who.int/en/health-topics/noncommunicable-diseases/mental-health/news/news/2012/10/depression-in-europe/depression-in-europe-facts-and-figures

[49] F. Chiarugi, et al., "Facial signs and psycho-physical status estimation for well-being assessment," in *Proc. 7th Int. Conf. Health Informat.*, 2014, pp. 555–562.

[50] M. Pediaditis, et al., "Extraction of facial features as indicators of stress and anxiety," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2015, pp. 3711–3714.

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

[52] M. Valstar, B. Schuller, J. Krajewski, R. Cowie, and M. Pantic, "Workshop summary for the 3rd international audio/visual emotion challenge and workshop (AVEC'13)," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 1085–1086.

[53] M. Valstar, et al., "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proc. 4th ACM Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 3–10.

[54] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, R. Cowie, and M. Pantic, "Summary for AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. ACM Multimedia Conf.*, 2016, pp. 1483–1484. [Online]. Available: http://doi.acm.org/10.1145/2964284.2980532

[55] M. Brown, A. Glendenning, E. A. Hoon, and A. John, "Effectiveness of web-delivered acceptance and commitment therapy in relation to mental health and well-being: A systematic review and meta-analysis," *J. Med. Internet Res.*, vol. 18, no. 8, Aug. 2016, Art. no. e221.

[56] M. Valstar, "Automatic behaviour understanding in medicine," in *Proc. Workshop Roadmapping Future Multimodal Interaction Res.*, 2014, pp. 57–60.

[57] N. Clark, T. Herman, J. Halverson, and H. K. Trivedi, "Technology tools supportive of DSM-5: An overview," in *Mental Health Practice in a Digital World: A Clinicians Guide*. Cham, Switzerland: Springer, 2015, pp. 199–211.

[58] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, U.K., Keele Univ.*, vol. 33, no. 2004, pp. 1–26, 2004, http://www.ifs.tuwien.ac.at/~weippl/systemicReviewsSoftwareEngineering.pdf

[59] N. Jähne-raden, C. Scharnweber, and R. Haux, "Utilization of health-enabling technologies for early detection of symptom parameters of addiction and depression," in *Proc. 24th Int. Conf. Eur. Federation Med. Informat. Quality Life Through Quality Inf.*, 2012, http://person.hst.aau.dk/ska/mie2012/CD/Interface_MIE2012/MIE_2012_Content/MIE_2012_Content/sco.html

[60] S. K. D'Mello, "Automated mental state detection for mental health care," in *Artificial Intelligence in Behavioral and Mental Health Care*. Cambridge, MA, USA: Academic Press, 2015, pp. 117–136.

[61] B. W. Schuller, "Acquisition of affect," in *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*. Cham, Switzerland: Springer, 2016, pp. 57–80.

[62] B. Martinez and M. F. Valstar, "Advances, challenges, and opportunities in automatic facial expression recognition," in *Advances in Face Detection and Facial Image Analysis*. Cham, Switzerland: Springer, 2016, pp. 63–100.

[63] M. P. Hyett, G. B. Parker, and A. Dhall, "The utility of facial analysis algorithms in detecting melancholia," in *Advances in Face Detection and Facial Image Analysis*. Cham, Switzerland: Springer, 2016, pp. 359–375.

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

[65] C. A. Corneanu, M. O. Simn, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016.

[66] H. Ellgring, *Non-Verbal Communication in Depression*, R. J. Eiser and K. R. Scherer, Eds. New York, NY, USA: Cambridge Univ. Press, 2007.

[67] P. H. Waxer, "Therapist training in nonverbal communication. I: Nonverbal cues for depression," *J. Clinical Psychology*, vol. 30, no. 2, 1974, Art. no. 215.

[68] B. Hosseinifard, M. H. Moradi, and R. Rostami, "Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal," *Comput. Methods Programs Biomed.*, vol. 109, no. 3, pp. 339–345, 2013.

[69] R. Adorni, A. Gatti, A. Brugnera, K. Sakatani, and A. Compare, "Could fNIRS promote neuroscience approach in clinical psychology?" *Frontiers Psychology*, vol. 7, 2016, Art. no. 456. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00456

[70] T. Suto, M. Fukuda, M. Ito, T. Uehara, and M. Mikuni, "Multichannel near-infrared spectroscopy in depression and schizophrenia: Cognitive brain activation study," *Biol. Psychiatry*, vol. 55, no. 5, pp. 501–511, 2004.

[71] G. J. Siegle, S. R. Steinhauer, E. S. Friedman, W. S. Thompson, and M. E. Thase, "Remission prognosis for cognitive therapy for recurrent depression using the pupil: Utility and neural correlates," *Biol. Psychiatry*, vol. 69, no. 8, pp. 726–733, 2011.

[72] J. S. Silk, et al., "Pupillary reactivity to emotional information in child and adolescent depression: Links to clinical and ecological measures," *Amer. J. Psychiatry*, vol. 164, no. 12, pp. 1873–1880, 2007.

[73] N. P. Jones, G. J. Siegle, and D. Mandell, "Motivational and emotional influences on cognitive control in depression: A pupillometry study," *Cogn. Affective Behavioral Neurosci.*, vol. 15, no. 2, pp. 263–275, 2014.

[74] J. Wang, Y. Fan, X. Zhao, and N. Chen, "Pupillometry in Chinese female patients with depression: A pilot study," *Int. J. Environ. Res. Public Health*, vol. 11, no. 2, pp. 2236–2243, 2014.

[75] D. Zhou, et al., "Tackling mental health by integrating unobtrusive multimodal sensing," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1401–1408.

[76] A. Y. Kudinova, K. L. Burkhouse, G. Siegle, M. Owens, M. L. Woody, and B. E. Gibb, "Pupillary reactivity to negative stimuli prospectively predicts recurrence of major depressive disorder in women," *Psychophysiology*, 2016. [Online]. Available: http://dx.doi.org/10.1111/psyp.12764

[77] M. Li, S. Lu, G. Wang, L. Feng, B. Fu, and N. Zhong, "Alleviated negative rather than positive attentional bias in patients with depression in remission: An eye-tracking study," *J. Int. Med. Res.*, vol. 44, pp. 1072–1086, 2016.

[78] R. B. Price, et al., "From anxious youth to depressed adolescents: Prospective prediction of 2-year depression symptoms via attentional bias measures," *J. Abnormal Psychology*, vol. 125, no. 2, pp. 267–278, 2015.

[79] C. Winograd-Gurvich, N. Georgiou-Karistianis, P. Fitzgerald, L. Millist, and O. White, "Ocular motor differences between melancholic and non-melancholic depression," *J. Affect. Disorders*, vol. 93, no. 1, pp. 193–203, 2006.

[80] C. Winograd-Gurvich, N. Georgiou-Karistianis, P. B. Fitzgerald, L. Millist, and O. B. White, "Self-paced and reprogrammed saccades: Differences between melancholic and non-melancholic depression," *Neurosci. Res.*, vol. 56, no. 3, pp. 253–260, 2006.

[81] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System: The Manual*. Salt Lake City, UT, USA: Res. Nexus Netw. Inf. Res., 2002.

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

[83] G. McIntyre, R. Göcke, M. Hyett, M. Green, and M. Breakspear, "An approach for automatically measuring facial activity in depressed subjects," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interaction Workshops*, 2009, pp. 1–8.

[84] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and PTSD: Exploring gender differences," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interaction*, 2013, pp. 147–152.

[85] S. Ghosh, M. Chatterjee, and L.-P. Morency, "A multimodal context-based approach for distress assessment," in *Proc. 16th Int. Conf. Multimodal Interaction*, 2014, pp. 240–246.

[86] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and PTSD: The effect of gender," *J. Multimodal User Interfaces*, vol. 9, no. 1, pp. 17–29, 2014.

[87] G. J. McIntyre, "The computer analysis of facial expressions: On the example of depression and anxiety," Ph.D. dissertation, College Eng. Comput. Sci., Australian Nat. Univ., Canberra, Australia, 2010.

[88] J. F. Cohn, "Social signal processing in depression," in *Proc. 2nd Int. Workshop Social Signal Process.*, 2010, pp. 1–2.

[89] G. Mcintyre, R. Göcke, M. Breakspear, and G. Parker, "Facial response to video content in depression," in *Proc. 13th Int. Conf. Multimodal Interaction Workshop: Inferring Cogn. Emotional States Multimodal Measures*, 2011, https://icmi.acm.org/2011/data/ICMI_2011_Technical_Program.pdf

[90] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image Vis. Comput.*, vol. 32, no. 10, pp. 641–647, 2013.

[91] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald, "Social risk and depression: Evidence from manual and automatic facial expression analysis," in *Proc. 10th Int. Conf. Autom. Face Gesture Recognit.*, 2013, pp. 1–8.

[92] J. F. Cohn, "Beyond group differences: Specificity of nonverbal behavior and interpersonal communication to depression severity," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, 2013, pp. 1–2.

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

[94] S. Poria, A. Mondal, and P. Mukhopadhyay, "Evaluation of the intricacies of emotional facial expression of psychiatric patients using computational models," in *Understanding Facial Expressions in Communication*. Berlin, Germany: Springer, 2015, pp. 199–226.

[95] T.-H. Yang, C.-H. Wu, K.-Y. Huang, and M.-H. Su, "Coupled HMM-based multimodal fusion for mood disorder detection through elicited audio–visual signals," *J. Ambient Intell. Humanized Comput.*, pp. 1–12, 2016, https://link.springer.com/article/10.1007/s12652-016-0395-y#citeas

[96] M. Valstar, et al., "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 3–10. [Online]. Available: http://doi.acm.org/10.1145/2988257.2988258

[97] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*. Hoboken, NJ, USA: Wiley, 2005, pp. 45–60.

[98] S. Scherer, et al., "Automatic behavior descriptors for psychological disorder analysis," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.

[99] Y. Zhou, et al., "Multimodal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs," in *Proc. 17th Workshop Semantics Pragmatics Dialogue*, 2013, pp. 160–169.

[100] L.-P. Morency, et al., "SimSensei demonstration: A perceptive virtual human interviewer for healthcare applications," in *Proc. 29th Conf. Artif. Intell.*, 2015, pp. 4307–4308.

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

[102] J. Gratch, et al., "The distress analysis interview corpus of human and computer interviews," in *Proc. Language Resources Eval. Conf.*, 2014, pp. 3123–3128.

[103] S. Scherer, et al., "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image Vis. Comput.*, vol. 32, no. 10, pp. 648–658, 2014.

[104] G. M. Lucas, J. Gratch, S. Scherer, J. Boberg, and G. Stratou, "Towards an affective interface for assessment of psychological distress," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2015, pp. 539–545.

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

[106] S. Alghowinem, R. Göcke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear, "Cross-cultural detection of depression from nonverbal behaviour," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2015, pp. 1–8.

[107] R. Gupta, et al., "Multimodal prediction of affective dimensions and depression in human-computer interactions categories and subject descriptors," in *Proc. 4th ACM Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 33–40.

[108] S. Alghowinem, R. Göcke, M. Wagner, G. Parker, and M. Breakspear, "Head pose and movement analysis as an indicator of depression," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interaction*, 2013, pp. 283–288.

[109] J. Joshi, "Depression analysis: A multimodal approach," in *Proc. 14th ACM Int. Conf. Multimodal Interaction*, 2012, pp. 321–324.

[110] J. Joshi, "An automated framework for depression analysis," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interaction*, 2013, pp. 630–635.

[111] J. Joshi, R. Göcke, G. Parker, and M. Breakspear, "Can body expressions contribute to automatic depression analysis?" in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–7.

[112] J. Joshi, A. Dhall, R. Göcke, M. Breakspear, and G. Parker, "Neural-net classification for spatio-temporal descriptor based depression analysis," in *Proc. 21st Int. Conf. Pattern Recognit.*, 2012, pp. 2634–2638.

[113] J. Joshi, A. Dhall, R. Göcke, and J. F. Cohn, "Relative body parts movement for automatic depression analysis," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interaction*, 2013, pp. 492–497.

[114] A. Pampouchidou, K. Marias, M. Tsiknakis, P. Simos, F. Yang, and M. Fabrice, "Designing a framework for assisting depression severity assessment from facial image analysis," in *Proc. IEEE Int. Conf. Signal Image Process. Appl.*, 2015, pp. 578–583.

[115] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed. Lake St. Louis, MO, USA: American Psychiatric Assoc., 1994.

[116] S. Alghowinem, et al., "Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors," *IEEE Trans. Affect. Comput.*, 2016, doi: 10.1109/TAFFC.2016.2634527.

[117] B. Ooi KuanEe, L.-S. A. Low, M. Lech, and N. Allen, "Prediction of clinical depression in adolescents using facial image analaysis," in *Proc. 12th Int. Workshop Image Anal. Multimedia Interactive Serv.*, 2011, http://toc.proceedings.com/21884webtoc.pdf

[118] J. A. Coan and J. J. Allen, *Handbook of Emotion Elicitation and Assessment*. London, U.K.: Oxford Univ. Press, 2007.

[119] A. J. Fridlund, *Human Facial Expression: An Evolutionary View*. Cambridge, MA, USA: Academic Press, 1994.

[120] A. J. Fridlund, The Behavioral Ecology View of Facial Displays, 25 Years Later, Aug. 2015. [Online]. Available: http://emotionresearcher.com/the-behavioral-ecology-view-of-facial-displays-25-years-later/

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

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

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

[124] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic depression scale prediction using facial expression dynamics and regression," in *Proc. 4th ACM Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 73–80.

[125] B. Ooi Kuan Ee, "Early prediction of clinical depression in adolescents using single-chanel and multichannel classification approach," Ph.D. dissertation, School Elect. Comput. Eng., RMIT Univ., Melbourne, VIC, Australia, 2014.

[126] M. Sidorov and W. Minker, "Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach," in *Proc. 4th ACM Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 81–86.

[127] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker, "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression," in *Proc. 3rd Int. Conf. Pattern Recognit. Appl. Methods*, 2014, pp. 671–678.

[128] M. Kächele, M. Schels, and F. Schwenker, "Inferring depression and affect from application dependent meta knowledge," in *Proc. 4th ACM Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 41–48.

[129] M. Senoussaoui, M. Sarria-Paja, J. A. F. Santos, and T. H. Falk, "Model fusion for multimodal depression classification and level detection," in *Proc. 4th ACM Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 57–63.

[130] I. T. Meftah, N. L. Thanh, and C. B. Amar, "Detecting depression using multimodal approach of emotion recognition," in *Proc. Int. Conf. Complex Syst.*, 2012, pp. 1–6.

[131] Y. Katryal, S. V. Alur, S. Dwivedi, and R. Menaka, "EEG signal and video analysis based depression indication," in *Proc. Int. Conf. Adv. Commun. Control Comput. Technol.*, 2014, pp. 1353–1360.

[132] J. J. Maller, S. S. George, R. P. Viswanathan, P. B. Fitzgerald, and P. Junor, "Using thermographic cameras to investigate eye temperature and clinical severity in depression," *J. Biomed. Optics*, vol. 21, no. 2, 2016, Art. no. 026001.

[133] S. Harati, A. Crowell, H. Mayberg, J. Kong, and S. Nemati, "Discriminating clinical phases of recovery from major depressive disorder using the dynamics of facial expression," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2016, pp. 2254–2257.

[134] S. Alghowinem, "Multimodal analysis of verbal and nonverbal behaviour on the example of clinical depression," Ph.D. dissertation, Australian Nat. Univ., College of Engineering and Computer Science, Canberra, ACT, Australia, 2015.

[135] H. Pérez Espinoza, H. J. Escalante, L. Villaseñor Pineda, M. Montes-y Gómez, D. Pinto-Avedaño, and V. Reyes-Meza, "Fusing affective dimensions and audio-visual features from segmented video for depression recognition," in *Proc. 4th ACM Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 49–55.

[136] H. Dibeklioğlu, Z. Hammal, Y. Yang, and J. F. Cohn, "Multimodal detection of depression in clinical interviews," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 307–310.

[137] C. Smailis, N. Sarafianos, T. Giannakopoulos, and S. Perantonis, "Fusing active orientation models and mid-term audio features for automatic depression estimation," in *Proc. 9th ACM Int. Conf. PErvasive Technol. Related Assistive Environ.*, 2016, Art. no. 39.

[138] N. C. Maddage, R. Senaratne, L.-S. A. Low, M. Lech, and N. Allen, "Video-based detection of the clinical depression in adolescents," in *Proc. 31st Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2009, pp. 3723–3726.

[139] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression estimation using audiovisual features and fisher vector encoding," in *Proc. 4th ACM Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 87–91.

[140] S. Bhatia, "Multimodal sensing of affect intensity," in *Proc. 18th ACM Int. Conf. Multimodal Interaction*, 2016, pp. 567–571. [Online]. Available: http://doi.acm.org/10.1145/2993148.2997622

[141] S. Bhatia, M. Hayat, M. Breakspear, G. Parker, and R. Goecke, "A video-based facial behaviour analysis approach to melancholia," in *Proc. 12th IEEE Conf. Autom. Face Gesture Recognit.*, 2017, pp. 754–761.

[142] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble CCA for continuous emotion prediction," in *Proc. 4th ACM Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 19–26.

[143] A. Dhall and R. Goecke, "A temporally piece-wise Fisher vector approach for depression analysis," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2015, pp. 255–259.

[144] A. Pampouchidou, et al., "Video-based depression detection using local curvelet binary patterns in pairwise orthogonal planes," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2016, pp. 3835–3838.

[145] H. Kaya and A. A. Salah, "Eyes whisper depression: A CCA based multimodal approach," in *Proc. Int. Conf. Multimedia*, 2014, pp. 961–964.

[146] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, "Decision tree based depression classification from audio video and language information," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 89–96. [Online]. Available: http://doi.acm.org/10.1145/2988257.2988269

[147] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 7, pp. 1432–1441, Jul. 2015.

[148] L. He, D. Jiang, and H. Sahli, "Multimodal depression recognition with dynamic visual and audio cues," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2015, pp. 260–266.

[149] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Trans. Affect. Comput.*, 2017, doi: 10.1109/TAFFC.2017.2650899.

[150] A. Pampouchidou, et al., "Depression assessment by fusing high and low level features from audio, video, and text," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 27–34. [Online]. Available: http://doi.acm.org/10.1145/2988257.2988266

[151] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 43–50. [Online]. Available: http://doi.acm.org/10.1145/2988257.2988261

[152] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders," *J. Neurosci. Methods*, vol. 200, no. 2, pp. 237–256, 2011.

[153] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[154] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.

[155] M. Schröder, "The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems," *Advances Human-Comput. Interaction*, vol. 2010, 2010, Art. no. 2.

[156] G. Littlewort, et al., "The computer expression recognition toolbox (CERT)," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, 2011, pp. 298–305.

[157] K. A. Funes Mora, L. Nguyen, D. Gatica-Perez, and J.-M. Odobez, "A semi-automated system for accurate gaze coding in natural dyadic interactions," in *Proc. 15th ACM Int. Conf. Multimodal Interaction*, 2013, pp. 87–90.

[158] K. A. Funes-Mora and J.-M. Odobez, "Gaze estimation in the 3D space using RGB-D sensors," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 194–216, 2016.

[159] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3D face alignment from 2D videos in real-time," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, May 2015, pp. 1–8.

[160] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 1034–1041.

[161] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System: Investigators Guide*. Salt Lake City, UT, USA: Res. Nexus Netw. Inf. Res., 2002.

[162] G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau, "Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A Rreview," *Comput. Biol. Med.*, vol. 60, pp. 8–31, 2015.

[163] P. Wang, et al., "Automated video-based facial expression analysis of neuropsychiatric disorders," *J. Neurosci. Methods*, vol. 168, no. 1, pp. 224–238, 2008.

[164] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE J. Biomed. Health Informat.*, 2017, doi: 10.1109/JBHI.2017.2676878.

[165] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: Machine learning and application to spontaneous behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 568–573.

[166] S. Lucey, I. Matthews, Z. Ambadar, J. Cohn, F. de la Torre, and C. Hu, "AAM derived face representations for robust facial action recognition," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2006, pp. 155–162.

[167] J. R. Williamson, et al., "Detecting depression using vocal, facial and semantic communication cues," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 11–18. [Online]. Available: http://doi.acm.org/10.1145/2988257.2988263

[168] Z. Huang, et al., "Staircase regression in OA RVM, data selection and gender dependency in AVEC 2016, " in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 19–26. [Online]. Available: http://doi.acm.org/10.1145/2988257.2988265

[169] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, 1968, Art. no. 213.

[170] F. Morbini, D. Devault, K. Georgila, R. Artstein, D. Traum, and L.-P. Morency, "A demonstration of dialogue processing in SimSensei Kiosk," in *Proc. 15th Annu. Meeting Special Interest Group Discourse Dialogue*, 2014, Art. no. 254.

[171] D. DeVault, et al., "SimSensei Kiosk: A virtual human interviewer for healthcare decision support," in *Proc. Int. Conf. Auton. Agents Multi-Agent Syst.*, 2014, pp. 1061–1068.

[172] G. Stratou and L.-P. Morency, "MultiSense—Context-aware nonverbal behavior analysis framework: A psychological distress use case," *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 190–203, Apr.–Jun. 2017.

[173] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, 2013, pp. 41–48.

[174] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge 2013, " in *Proc. 15th ACM Int. Conf. Multimodal Interaction*, 2013, pp. 509–516.

[175] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *Proc. 16th Int. Conf. Multimodal Interaction*, 2014, pp. 461–466.

[176] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: EmotiW 2015, " in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 423–426.

[177] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001.

Anastasia Pampouchidou received the bachelor's degree in applied informatics and multimedia, from the Technological Educational Institute of Crete, in 2004 and the master's degree in computer vision from the University of Burgundy, in 2011, where she is currently working toward the PhD degree since 2013. Her PhD studies are funded by the Greek State Scholarships Foundation, under the scholarship programme from the revenue of the legacy in memory of Mary Zaousi. Previously, she worked for the EU funded project SEMEOTICONS for a period of two years. Her research interests involve image processing, machine learning, facial expression analysis, and affective computing.

Panagiotis G. Simos received the PhD degree in experimental psychology-biopsychology from Southern Illinois University, in 1995. He served as assistant and associate professor with the Departments of Neurosurgery, University of Texas Houston Medical School, and Psychology, University of Crete, Greece. He is currently professor of developmental neuropsychology in the School of Medicine, University of Crete. His research has been supported by several federal and national grants and focuses on neuropsychological and brain imaging studies of reading and memory using Magnetoencephalography and MRI/fMRI with children and adults. Ongoing studies explore psychoeducational, emotional, and neurophysiological profiles associated with specific reading disability, ADHD and neurodegenerative disorders. He has also developed and adapted in Greek several psychometric instruments for cognitive and linguistic abilities across the life-span.

Kostas Marias is an associate professor in image processing in the Informatics Engineering Department, Technological Educational Institute of Crete and since 2010 he is the head and founder of the Computational Biomedicine Laboratory, FORTH-ICS (previously Biomedical Informatics Laboratory). Previously, he was a principal researcher in the Institute of Computer Science (ICS-FORTH) since 2006. During 2000-2002, he worked as a researcher with the University of Oxford and from 2003-2006 as associated researcher at FORTH-ICS. He was the coordinator two EC projects on cancer modelling (http://www.contracancrum.eu/ and http://www.tumor-project.eu/), while during 2010-2015 actively participated in several other EC funded projects developing ICT technology focusing on medical image processing and personalized medicine. He has published more than 150 papers in international journals, books, and conference proceedings focusing on medical image analysis, biomedical informatics and modelling for personalized medicine.

Fabrice Meriaudeau received both the master's degree in physics from Dijon University, France as well as an engineering degree (FIRST) in material sciences, in 1994. He also received the PhD degree in image processing with the Dijon University, in 1997. He was a postdoc for a year with The Oak Ridge National Laboratory. He is currently "professeur des Universites" and was director of the Le2i (UMR CNRS), which has more than 200 staff members, from 2011 to 2016. His research interests were focused on image processing for non-conventional imaging systems (UV, IR, polarization) and more recently on medical/biomedical imaging. He coordinated an Erasmus Mundus Master in the field of Computer Vision and Robotics from 2006 to 2010 and he was the vice president for International Affairs for the University of Burgundy from 2010 to 2012. He has authored and co-authored more than 150 international publications and holds three patents. Since 2016, he is on leave from "Université de Bourgogne" and has joined Universiti Teknologi PETRONAS as a professor.

Fan Yang received the BS degree in electrical engineering from the University of Lanzhou, China, in 1982 and the MS (computer science) and PhD degrees (image processing) from the University of Burgundy, France, in 1994 and 1998, respectively. She is currently a full professor and member of LE2I UMR CNRS, Laboratory of Electronic, Computing, and Imaging Sciences, University of Burgundy, France. Her research interests include the pattern recognition, neural network, parallelism and real time implementation, and, more specifically, automatic face image processing algorithms and architectures. She is member of the French research group ISIS (Information, Signal, Images and Vision), she livens up the theme C: Algorithm Architecture Mapping.

Matthew Pediaditis received the engineer's and PhD degrees in biomedical engineering from Graz University of Technology, Austria. He has collaborated in numerous research projects in the domains of eHealth, ambient assisted living and patient health monitoring. He has expertise in video analysis for health applications using computer vision and machine learning. His current research focuses on human motion and facial expression analysis for disease assessment in epilepsy.

Manolis Tsiknakis received the BEng degree in electric and electronic engineering, in 1983, the MSc degree in microprocessor engineering, in 1985, and the PhD degree in systems engineering from the University of Bradford, Bradford, United Kingdom, in 1989. From February 1992 until January 2012 he has been with the Institute of Computer Science, Foundation for Research and Technology-Hellas, Greece, as a principal researcher and head of the center of eHealth Technologies. He is currently a professor of biomedical informatics in the Department of Informatics Engineering, Technological Educational Institute of Crete and a visiting researcher at FORTH/ICS. He led the effort for the design and implementation of HYGEIAnet, the regional health informatics network of Crete, an eEurope award winner in 2002 and was FORTH's best applied research award winner in 2003. He is currently an associate editor in the *IEEE Journal of Biomedical and Health Informatics*. He has published extensively - more than 250 papers - in refereed scientific journals and international conferences on issues related to the application of innovative ICT in the domain of clinical and translational research, care and wellness management. His current research interests include biomedical informatics and engineering, approaches for semantic health data integration, service oriented SW architectures and cloud computing and their application in biomedicine, affective computing, behavioral modeling and human activity recognition using wearable sensors, smart eHealth and mHealth service platforms.