Automatic Diagnosis of Mental Disorders

By

Sarmad AL-gawwam

A doctoral thesis submitted in fulfilment of the requirements for the award of Doctor of Philosophy

The University of Sheffield
Department of Electrical and Electronic Engineering

January 2021

Supervisor
Dr. Mohammed Benaissa

Automatic Diagnosis of Mood Disorders

Sarmad AL-gawwam

PhD Thesis

Communications Group
Department of Electrical and Electronic Engineering
The University of Sheffield

2021

This thesis is dedicated my beloved parents, for their continuous love and encouragement and my wife for her endless support, patience and believe in me.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Dr. Mohammed Benaissa. for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides, I am grateful Dr. Aleksandr Zaitcev for giving the encouragement and sharing insightful suggestions that allowed me to do the most valuable part of my research. This part of my thesis would not be possible without him.

This thesis would not have been possible without the endless love and support of my family who have always encouraged me to follow my passion. I am grateful to them for their unconditional love and support whenever I needed.

Abstract

A significant number of people worldwide suffer from mental disorders such as depression, bipolar and neurodegenerative disorders. These disorders adversely impact the life quality of people and have a significant economic impact on health-care providers. The World Health Organisation (WHO) considers depression as a common mental disorder, with more than 300 million people of all ages affected by depression. Similarly, Bipolar disorder affects more than 60 million individuals, which makes it among the most spread mental disorders worldwide. Even though the treatment of mental disorders has proved to be efficient in most situations, incorrect diagnosis is a common obstacle. This is because self-administered questionnaires and clinical interviews are the only available methods for diagnosis. These methods are influenced by subjective bias from clinicians or patients, time-consuming and hard to repeat. There is growing attention on early diagnosis of mental disorders as evolving treatments are expected to be more effective before irrevocable changes have occurred in the brain. The integration of novel methods based on the automatic analysis of visual signals may provide more information about a person's mental state, which could contribute to the clinical diagnostic process.

This thesis demonstrates that eye features extracted from video recordings of patients' answers to a clinician's questions can help in the automated diagnosis of depression. Results show that there is eye blink abnormality among depressed patients due to psychomotor retardation. This manifests as longer eye blink duration. The efficacy of these features demonstrated in depression severity prediction where it achieved mean absolute error (MAE) of 8.30 and classification accuracy of 93% using the Audio/Visual Emotion Challenge 2014 (AVEC2014) dataset. Furthermore, visual features of eye movements combined with head pose features extracted from video recordings of patient's during clinical interview in a specialist memory clinic for the development of a automated visual screening method to support the preliminary detection of patients with cognitive concerns related to progressive neurodegenerative disorders (ND), Functional Memory Disorder (FMD) and Mild Cognitive Impairment (MCI).

Finally, a novel automatic diagnosis method for bipolar disorder developed based on deep learning. The proposed method investigated the application of several deep neural network architectures to extract the complex temporal trajectories of physiological behaviours that occur at multiple time scales. Results achieved shows that this model can be used as a universal automatic feature extractor for mental disorders. This is confirmed by testing the proposed model on AVEC2018 bipolar disorder dataset and AVEC2014 depression dataset. In the AVEC2018 dataset, individuals with bipolar disorder are classified into states of remission, hypo-mania and mania. The proposed model achieved Unweighted Average Recall (UAR) of 55.56%. Using the same model on AVEC2014 depression dataset, Mean Absolute Error (MAE) of 7.65 achieved which outperforms most of the previous studies on the same dataset. These results confirm the generalisability of the proposed deep learning model and that it can be used as a tool for multi-scale feature extraction.

Contents

List of Figures

List of Tables

List of Publications

Journal papers

Al-gawwam, S., Benaissa, M. (2018). Robust eye blink detection based on eye landmarks and Savitzky–Golay filtering. Information, 9(4), 93.

Conference Papers

Al-gawwam, S., Benaissa, M. (2018, December). Depression Detection From Eye Blink Features. In 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (pp. 388-392). IEEE.

Al-Gawwam, S., Benaissa, M. (2017, December). Eye blink detection using facial features tracker. In Proceedings of the International Conference on Bioinformatics Research and Applications 2017 (pp. 27-30). ACM.

List of Abbreviations

AAM	Active Appearance Model.
AD	Alzheimers Disease.
ADHD	Attention Deficit Hyperactivity disorder.
ANN	Artificial Neural Network.
APA	American Psychiatric Association.
ASD	Autism Spectrum Disorder.
AUs	Action Units.
AVEC	Audio/Visual Emotion Challenge.
BDI	Beck Depression Inventory.
BN	Batch Normalisation.
CCA	Canonical Correlation Analysis.
CLM	Constrained Local Model.
CNN	Convolutional Neural Networks.
Conv1D	1 Diemension convolutional.
Conv2D	2 Diemension convolutional.
CT	Computerised Tomography.
DALY	Disability-Adjusted Life Year.
DNN	Deep Neural Netwotk.
DSC	Depression prediction Challenge.

DSM-5	Diagnostic and Statistical Manual of Mental Disorders.
ET	Extra Trees Classifier.
FACS	Facial Action Coding System.
FER	Facial expression recognition.
FMD	Functional Memory Disorders.
FNNS	Feed-forward neural networks.
FSM	Finite State Machine.
FWHM	Full Width at Half Maximum.
HCI	Human-Computer Interaction.
HDR	Histogram of Displacement Range.
HLDs	higher level descriptors.
HOG	Histogram of Oriented Gradients.
LBGP	Local Binary Gabor Patterns.
LBP	Local Binary Pattern.
LGBP-TOP	Local Gabor Binary Patterns from Three Orthogonal Planes.
LPC	Linear predictive coding.
LPQ	Extra Trees Classifier.
LSTM	Long Short Term Memory.
MAE	Mean Absolute Error.
MCI	Mild Cognitive Impairment.
MDD	Major Depressive Disorder.
MFCC	Mel-Frequency Cepstral Coefficients.
MLP	MultiLayer Perceptron.

MRI	Magnetic Resonance Imaging.
ND	Neurodegenerative Disorders.
NLP	Natural Language Processing.
PCA	Principal Component Analysis.
PHQ	Patient Health Questionnaire.
ReLu	Rectified Linear Unit.
REM	Rapid Eye Movement.
ResNets	Residual Networks.
RF	Random Forest.
RMSE	Root Mean Square Error.
SBS	sequential backward search.
SFS	sequential forward search.
SG	Savitzky Golay.
SVM	Support Vector Machine.
TLCNN	Long-term Convolutional Neural Network.
UAR	Unweighted Average Recall.
VGG	Visual Geometry Group.
WGD	Weighted Gradient Descriptor.
WHO	World Health Organisation.
YMRS	Young Mania Rating Scale.

Chapter 1

Introduction

Humans express their behaviour by a set of visual communicative signals that represent an individual's thoughts, feelings, and provide an understanding of a person's psyche. For this reason, the analysis of human behaviour can be useful for applications in the domains of mental health, psychology and computer interaction. For example, mental disorders like bipolar and depression directly affect a person's behaviour in daily life. Depression disorder is one of the common causes of disability and seriously affects human mental health in all age groups [1]. This disorder can highly affect an individual's behaviour, thoughts, feelings and capability to work and ranges in severity from mild, short episodes of sadness to severe, persistent depression [2]. Depression is a psychiatric mental disorder resulting from a sudden stressful event affecting an individual's life. This causes a continuous feeling of sadness, negativity and makes it difficult to do everyday responsibilities. According to the World Health Organisation (WHO), depression is the fourth most reason for disability worldwide and is expected to become the second in 2020 due to its growing prevalence [3]. In the UK, depression affects 2.69 million people (4.5% of the total population) from different ages [4]. Moreover, the annual estimation cost for depression in the UK ranges between £7-£9 billion including the expenditures of preventive treatment, medical treatment and care [5, 6]. While in the United States, people with depression reported at 17.49 million (5.9% of the total population) [4]. Severe depression cases affecting not only the brain but the heart also, therefore increases the risk for a range of medical conditions like; cardiovascular disease, Alzheimer's disease (AD), vascular dementia, cancer, and stroke [7, 8]. In the same way, Bipolar Disorder (BD) is characterised by recurring episodes of depression (feelings of low mood and lethargy), and of mania (feelings of elation and over-activity) [9]. Typically, manic episodes include hyperactivity, irritable mood, loud speech and a decreased period of sleep [10]. The World Health Organisation (WHO) estimates that BD impacts more than 60 million individuals worldwide [11]. These mental disorders have placed an excessive burden on individuals, families and health care services. For example, the health services' annual expenditure due to the bipolar disorder in England is estimated to be £8.21 billion by 2026 [5]. As BD is a life-long illness, early diagnosis and following treatment can positively impact individuals' life quality with this disorder.

Neurodegenerative disorders (ND) is another form of mental disorders, and it is leading to dementia and affects a person's visual behaviour. Increasing demand on earlier diagnosis of ND as new treatments is probably more effective before irreversible changes have occurred in the brain. There has been a huge increase of referrals of people with memory complaints from primary care to secondary care, resulting in significant pressure to diagnostic pathways [12]. The aim to find early diagnostic procedure has led to 600% increase in referrals to secondary memory clinics in the UK over the last 10 years [12]. Although these referrals have increased the identification of patients with ND, a large rate of the patients now referred to memory clinics have Functional Memory Disorder (FMD) and mild cognitive impairment (MCI) without but no evidence of cognitive deficit. This group of patients makes 50% of referrals to neurology-led secondary care memory services in the UK [13]. Currently, assessment methodologies for mental disorders rely on subjective patient self-report or clinical evaluation of symptoms severity using a range of assessment methods such as Beck Depression Inventory (BDI) [14], Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [15] and Young Mania Rating Scale (YMRS) [16]. These methods require a comprehensive assessment by experienced professionals and are susceptible to diagnosis bias, which affects the accurate diagnosis as the number of people who have a mental disorder is increasing [17]. Therefore, automated diagnosis methods are anticipated to help clinicians provide an objective assessment and time saving automated diagnosis. In this thesis, the visual appearance features for the diagnosis of depression, ND and BD were investigated. These mental disorders affect the patients' multiple visual signals, which clinicians may not observe due to human limitations.

For example, depressed patients are characterised with a variety of well-known symptoms such as sadness or low mood and slower overall facial muscle movements [18]. While, BD patients shows episodes of depression, aggressive behaviour and high mood. Also, changes in psychomotor activities which require coordination of the body and brain to function efficiently.

Mental disorders are apparent as an altered visual activity such as facial expressions, head pose, and eye movements [19–22]. For the study of eye blinks, previous researches employed person-specific models that must be trained for each subject to track the face and extract eye movement features. Clearly, this approach is not usable for clinical applications. To this end, this thesis proposes a novel algorithm for eye blinks detection based on an automatic person independent facial landmark detection approach. The proposed eye blink detection algorithm is then used to investigate eye features' performance for depression disorder detection. Furthermore, visual features from eye blink, head pose and eye gaze are utilised to develop a novel automated diagnosis system for ND. For BD research, a novel automated diagnosis approach based on deep learning developed for temporal modelling of facial micro-expression trajectories to diagnose mental disorders. The proposed deep learning architecture's performance was validated by obtaining state-of-the-art inference accuracy on the AVEC2014 and the AVEC2018 datasets for depression and bipolar disorder diagnosis. Therefore, the proposed deep learning architecture can be employed for the diagnosis of multiple mental disorder.

1.1 Problem description

The lack of a cure for mental disorders requires an early diagnosis of the condition. Earlier diagnosis can improve the quality of life and slow the disease's progress, especially before the brain's irreversible changes. Identifying patients with mental disorders at early stages is challenging due to the lack of valid predictive markers appropriate for routine screening. Biomarkers that can identify patients at high risk of developing progressive mental disorder are expensive and only available in few clinical centres (e.g. amyloid Positron Emission Tomography test) or expose people to radiation (e.g. amyloid and tau testing in the cerebrospinal fluid test). Therefore, mental disorder diagnosis depends on subjective assessment reports from patients, families, or clinicians. These assessments are usually available in multiple-choice questions administered by clinicians while the patient is interviewed or self-administered. For example, the Patient Health Questionnaire (PHQ-8) [23] is a commonly used self-report form for depression diagnosis quantified by the test scores. This diagnosis includes questions on the overall mental, appetite, physical energy, motivation, and other difficult factors to measure directly. Using this type of assessment is not straightforward because the given test scores are sensitive to the patient's ability to report their symptoms honestly and willingly. As a result, the diagnosis process is a time-consuming and includes a significant amount of clinical training, qualifications and experience to get satisfactory results. Despite the fact that there were various biological markers linked with mental disorder, for example, low serotonin levels [24], genetic abnormalities [25] and neurotransmitter dysfunction [26], no specific biomarker has been determined. Hence, the lack of an objective standard for mental disorder diagnosis prevents the efforts of clinical services. This risks providing optimal patient care, placing an excessive burden on health, and economic and social utilities. The recent advancements in human behaviour understanding have inspired researchers to investigate the applicability of data-driven methods that can model various human visual behaviours that correlate with depression, BD or a neurodegenerative disorder. Such methods have the potential to significantly enhance the healthcare systems due to the benefits it can offer. Some of these benefits include understanding complex visual features that correlate with the given mental disorder and provide an objective assessment. Additionally, advances in automated diagnosis enable monitoring changes in multiple visual modalities affected by mental disorders such as eye blinks, eye gaze, head movements and facial expressions. Developing such a tool for visual features analysis is affordable, non-invasive, and can be deployed remotely. Moreover, deploying an automatic diagnosis tool based on visual features in primary health care is more likely to be useful, given that clinicians cannot capture the complex changes in the characteristics of visual signals. Therefore, the research presented in this thesis explores a solution for an automated mental disorder screening tool based on the analysis of a person's visual features.

1.2 Aims and objectives

This research explores advanced techniques to develop automated diagnosis methods to address the limitations introduced in the previous section. Automated mental disorder diagnosis has a huge potential to improve healthcare by providing objective measurements, interpretability, and information processing power. Moreover, once a developed method's satisfactory performance is achieved, the method can be replicated as many as required and consistently supply reproducible results. To achieve this aim, multi-modal visual features which is a combination of eye activity, head movements, facial action units and facial emotions are used to develop an automated diagnosis approach. This thesis's research is particularly concerned with analysing behavioural patterns during the speech that can be utilised by machine learning algorithms to distinguish mental disorder patients from healthy controls. This thesis is concerned with the following research questions:

1. The feasibility of using facial landmark detectors for the development of eye blink detection algorithm.

2. The feasibility of using eye features to identify depression automatically.

3. The feasibility of developing an automated system using visual features for the detection of progressive Neurodegenerative Disorders (ND), Mild Cognitive Impairment (MCI) and Functional Memory Disorders (FMD).

4. The feasibility of employing behavioural micro-expression for BD, What is the importance of extracting visual features on different time scales for automatic diagnosis of BD and which deep learning network architecture has the potential of extracting useful behavioural patterns.

The research focuses on investigating visual signals and machine learning techniques to derive clinically beneficial information from visual appearance. Hence, the aims have been set to:

- Develop a robust eye blink detection method that can be deployed in automated methods for mental health diagnosis

- Develop a new visual-based system that can be used to analyse doctors-patients conversations to understand memory disorder related to symptomology.

- Extract visual features that capture the complex behavioural patterns that happen during the speech in the presence of BD.

- Develop a deep learning network model that models visual features at different time scales to capture complex audio/visual emotions at a very subtle time duration.

- Test the generalisability of the proposed deep learning model and whether it can be used as a diagnostic tool for different mental disorders.

1.3 Thesis contributions

This thesis provides the following contributions:

1. **Novel eye blink detection method.**

   Eye blink detection methods that used for mental disorder diagnosis need tight requirements on the setup to account for image resolution, facial motion dynamics and head orientation. This makes them unsuitable as a low cost and robust solution to a wide range of applications. To tackle these limitations, the proposed solution investigates the usage of robust real-time facial landmark detectors that track key points of the human face, including eye corners and eyelids. The proposed eye blink detection approach uses landmark detector to track eye openness and apply signal filtering to detect eye blinks and extract related features. This system was evaluated using video recordings from five standard datasets. The complete approach is presented in chapter 3.

2. **An automated depression diagnosis method based on eye blink features.** This work proposes a new approach to diagnosing depression by utilising eye blink features extracted using the new eye blink algorithm presented in chapter 3. This approach provides more accurate features extraction as validated by several datasets. The AVEC2014 dataset used to validate the proposed depression diagnosis method. This study outperformed the baseline results reported on the AVEC2014 and similar to complex modalities

that used both speech and video features. The achieved accuracy ranging between 88% and 93% for depression classification task and Mean Absolute Error (MAE) of 8.30 for the test set. This work is presented in chapter 4.

3. **A new diagnostic approach for the identification of patients with neurodegenerative cognitive complaints.**

   Neurodegenerative diseases causing dementia affects a person's visual features such as eye blinks, eye gaze and head pose. Currently, the assessments in memory clinics examine patient's verbal abilities using verbal recall, comprehension and word fluency. The automatic analysis of visual signals may supply useful information about a person's state, which could help in the diagnostic process. Therefore, a method utilises visual features extracted from video recordings of patients' answers to a neurologist's questions in a memory clinic can support the preliminary diagnosis of patients with cognitive concerns to progressive ND, MCI or FMD. The proposed method's performance suggests that it could be incorporated into the diagnostic process for patients with neurodegenerative disorder. This work is presented in Chapter 5.

4. **Automatic screening system for bipolar disorder.**

   In this work, the usefulness of audiovisual features was evaluated for estimating the severity of the BD. This is done by utilising the temporal trajectories during the speech to develop a data-driven deep network for temporal feature extraction. Features were extracted at different time scales to capture complex audio/visual emotions at a very subtle time duration. The proposed method was tested using the AVEC2018 dataset, which is the only publicly available dataset for patients suffering from bipolar conditions. The proposed method showed the effectiveness of predicting the three states of bipolar, i.e. remission, hypo-mania, and mania with an unweighted average recall of 64% and 55.7% for development test partitions, respectively. These results were higher than the baseline results for audio and video modalities. Additionally, the AVEC2014 dataset was used to test the proposed method's generalisability on different mental health datasets. This experiment has provided state-of-the-art results on the AVEC2014 dataset with a mean absolute error of 7.65. These experiments show that the proposed method has the potential to be an automatic universal feature extractor. This system is presented in chapter 5, and 6.

1.4 Thesis outline

The overall structure of the remaining chapters in this thesis is organised as follows:

- **Chapter 2** provides an overview of the current techniques used for facial expression recognition and emotion-based application. Also, the chapter introduces various hand-crafted feature descriptors and machine learning algorithms employed in the next chapters. Finally, a literature review of studies that used visual features for mental disorder diagnosis is presented.

- **Chapter 3** Describes the new eye blink detection technique utilising facial landmarks. This technique's details were presented together with the video datasets that utilised to report the results. The proposed novel technique is compared with the existing eye blink detection algorithms, which have used the same datasets for evaluation.

- **Chapter 4** investigates the application of the proposed eye blink method for depression classification and predicting depression severity using regression score on the AVEC2014 dataset. The proposed system performance compared against other studies from the literature. Additionally, a new method is proposed to capture neurodegenerative cognitive-related symptoms by analysing the patient's visual features. Finally, this method is compared with a more complex approach used as a combination of three features to produce a similar performance.

- **Chapter 5** Introduces an automated method for Bipolar disorder diagnosis. This method built using visual appearance features prepared as time window sequences input to the machine learning pipeline and evaluated using the AVEC2018 Bipolar dataset. In this chapter, experiments were used to explore the bipolar disorder dataset and prepare machine learning predictions to be utilised in chapter 6.

- **Chapter 6** presents the development of a novel automated for screening bipolar disorder based on the multi-scale temporal modelling of visual features. For this task, audio/visual modalities have been utilised to develop convolutional neural network architectures for automatic extraction of feature temporal trajectories during the speech. The proposed deep neural network architectures were evaluated on the AVEC2014 and the AVEC2018 datasets and demonstrated state-of-the-art performance. Finally, the system performance also compared to other studies from the literature.
- **Chapter 7** Includes the summary with conclusions and further research directions.

Chapter 2

Literature review

This thesis aims to develop automated screening methods that can recognise individuals depending on particular visual biomarkers that differentiate between individuals with mental disorders and healthy controls. Automated screening method in the field of Affective Computing/ Social Signal Processing (AC/SSP) refers to a computational framework that can process information communicated in social signals to recognise various features of human behaviour [27]. Typical work-flow for developing an automatic system for mental disorder assessment proposed by [28] is shown in Figure 2.1 below. The first stage is to recognise a set of relevant social signals for the given task. These signals need to be represented as a measurable quantity before machine learning algorithms can process it. This type of representation is called a feature. Features extracted directly from social signals are called low-level descriptors (LLDs). These type of features are raw and needs subsequent processing before they can be ready to be passed to machine learning algorithms. The next step contains further processing of LLD features to get a more suitable form of features that machine learning algorithms can work with. Features obtained from this stage are known as higher-level descriptors (HLDs). HLDs are obtained using various feature engineerings tasks such as statistical summaries, dimensionality reduction and feature concatenation. Moreover, statistical measures can be utilised to determine the discriminative power and usability of HLDs before passing them to the next stage. The final stage of the automated screening pipeline is to feed these features into machine learning algorithms trying to produce a suitable result according to the research aim, i.e., existence of mental disorder or severity level assessment of the symptoms (e.g., bipolar intensity according to YMRS scores [16] ranging between 0-33 and depression intensity according to Beck Depression Inventory BDI-II scores [29] ranging between 0-63). As an example, if the target variable is continuous, the regression machine learning algorithms are selected. On the other hand, a classifier is selected if the target variable is discrete (e.g., depressed vs not depressed or low vs high depression severity). It is important to mention is that feature engineering and machine learning stages typically overlap and are tuned until an acceptable performance is achieved. This chapter aims to introduce visual features used to develop automated screening methods, feature engineering mechanisms, machine learning algorithms and visual features in mental disorders.

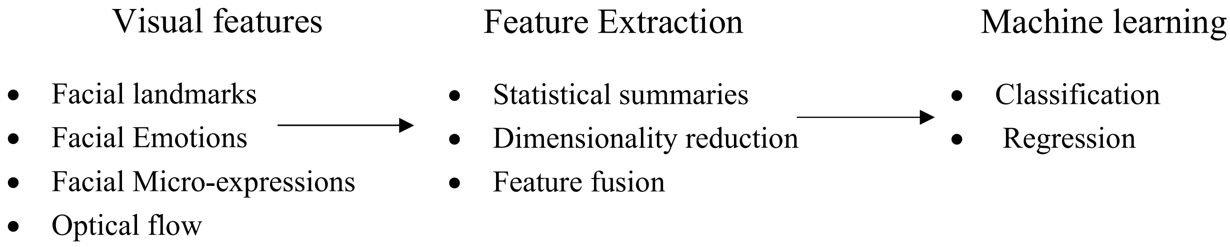


Figure 2.1 Typical work-flow for the development of an automatic system for mental disorder assessment.

2.1 Typical Facial Expression Recognition (FER) model

Automatic human visual appearance features analysis has been a hot topic for research. This is because of its usability in different applications such as human-computer interaction, mental health and bio-metrics, etc. A wide range of topics is included in visual features analysis, including detection of faces in images or video frames, localisation of landmarks, head pose estimation, facial emotion recognition and several others. The validity of these features are confirmed in research form psychology [30] [31]. This has inspired researchers to work on these features to develop automatic diagnosis systems for mental disorders such as schizophrenia [32] [33], depression [34] and autism [35] [36]. Other modalities such as hand gestures and leg movements might help automated screening; however, these features are not provided in most mental disorders datasets.

2.1.1 Face detection

The first step in any facial analysis system is face detection. This operation's output is usually shown as an output of a box highlighting the face region in an image or video. Due to variable lighting conditions, varied head poses and occlusions, face detection is still considered challenging. Viola and Jones [37] is the most popular algorithm for face detection in real-time. It employs haar-like features to train a cascade of Adaboost classifiers which is a time-efficient solution. However, this algorithm works well for near-frontal faces under normal conditions, and it becomes less efficient for faces under more wild conditions (lighting, expression and occlusion). Recently, deep learning has shown exceptional performance in object detection tasks [38] [39]. This has inspired many studies to improve face detection using the high capacity of deep convolutional networks. In [40], a CNN model was proposed to detect the face, pose and locations of facial landmarks. Other approaches were developed in [41] and [42].

2.1.2 Facial expressions

Facial expressions are one of the richest tools used to describe the nonverbal emotions conveying social information between humans. Individuals with mental disorders normally show altered emotional expressions that can be utilised to recognise them amongst healthy individuals potentially. Facial expression intensity can show clues about the person's intentions, physical health, emotions, pain or physical pleasure, and other physiological changes [43]. These expressions cause the movement of facial muscles corresponding to a certain emotion. As a result, these expressions make the corresponding facial muscles move and deliver a noticeable visual expression. The first study to note emotion recognition using a facial expression from psychology perspectives was done by [44]. According to [44], a set of seven basic emotions are apparent in humans beings regardless of culture, age and gender. The basic emotions are *Happy, Surprise, Neutral, Fear, Sad, Angry and Disgust*, as shown in Figure 2.2 [45]. These emotions have an observable structure when they occur; for example, anger emotion is characterised by eyebrows pulled down, upper lids pulled up, lower lids pulled up and tightened lips. Additionally, the work in [44] developed the Facial Action Coding System

(FACS) for describing relevant characteristics of facial muscle movement, as shown in Figure 2.3 [46]. Since then, it is widely used for facial expression because of the descriptive power that it provides for describing the details of facial expressions. There are 44 different Action Units (AUs), of which 33 are directly related to facial muscles' contraction, such as lifting the left eyebrow or the jaw tightening [46]. Although FACS provides an efficient technique for recognising facial expression details, it requires a professionally trained person to judge the expressions. This process could be time-consuming and prone to bias.

To overcome this limitation, many research efforts have been done for automatic detection of AUs form sequence or spatial-based data [47] [48]. There have been several public challenges introduced in [47, 48] encouraging researchers to come up with systems that tackle this problem, showing useful results in [49] and [50]. Recently, computationally modelling Major Depressive Disorder (MDD) using visual features such as facial expressions is an active research area. Depression is the most common mental disorder that affects a person's mental health and is represented by permanent negative feelings. According to [51], depression can be realised as a mental disability characterised by lack of interest, reduced energy, feelings of guilt and sleep disturbances. In a current report, the world health Organisation (WHO) estimated that 350 million people worldwide are affected by depression[52].

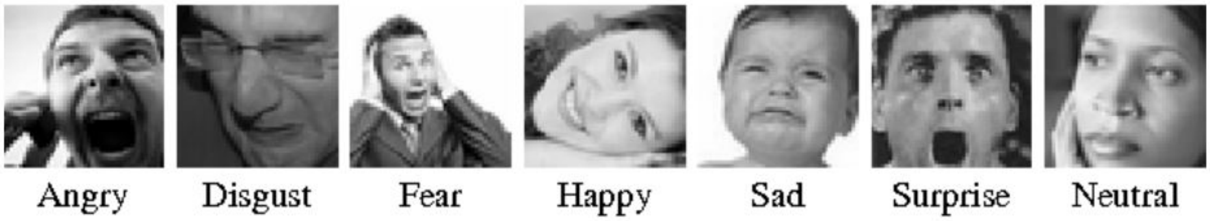
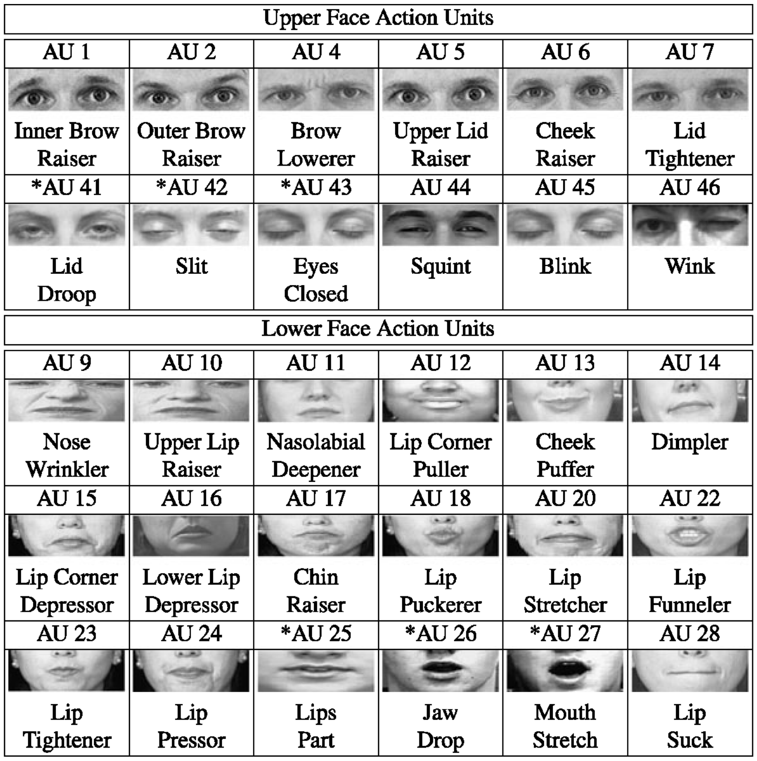


Figure 2.2 The seven basic emotions adopted from [45]

Moreover, depression is the fourth most significant cause of disability worldwide and is expected to be the leading cause in 2020 [53]. Currently, automated depression disorder diagnosis is an active research subject in the affective computing field to determine the nature and severity of depression in affected patients. There are several assessment methods for diagnosing depression, depending on the patient's review and reports during an interview. The severity level of depression evaluations varies depending on the clinician's experience and the methods used for diagnosis (e.g., Diagnostic and Statistical Manual of Mental Disorders (DSM-



Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink

Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Figure 2.3 Facial Action Units combinations [46].

IV) [54], PHQ-8 [23] and PHQ-9 [55]. Computer vision techniques have been utilised from these assessment methods to translate a patient's mental condition and facial representations into an objective depression scale. AUs used by [56] to analyse facial emotions rather than the seven basic emotions. They found that emotional expression depends highly on gender, where men have shown a higher rate of AU4 than women. This can be translated as men tend to smile under depression more than women. Also, the same trend of disgust has been observed for both genders with depression disorder. Lucas et al. [57] found that individuals with depression display facial expressions of hostility and reduced signs of joy. These observations agree with the psychomotor retardation reported in the literature from psychology.

2.1.3 Facial micro-expressions

Emotions are regarded as psycho-physiological experiences related to the individual's reaction to an event. Details from human visual interaction reveal an individual's feelings, social interaction and opinions. Besides the specific categories of emotions proposed by [44], emotions that are located between basic emotions are known as dimensional emotions.

According to the observations in [58], dimensional emotions are not expressed by individuals similarly and are believed to come from cultural or individual expressions of the basic emotions; therefore, dimensional emotions are closely related to human behaviour. Figure 2.4 [58] shows dimensional emotions as suggested in [58].

A clearly identified typical facial emotion expression lasts from 0.5 to 4 seconds [59]. However, many studies showed that depending on facial expression for emotion recognition may be tricky. This is because some people may try to suppress their genuine emotion by showing an opposite facial expression due to cultural and social reasons [60]. These expressions are known as micro-expressions, and they provide a significant amount of knowledge that can reveal depressed emotions [61] and understand the true mental condition of a person. By analysing depressed subjects video recordings, [62] observed intense expressions of anguish lasted only for two frames. This experiment highlighted the importance of micro-expression analysis for affect monitoring [63]. However, due to the concise period of occurrence of micro-expressions, which lasts for a very brief interval of time (i.e., 1/25s to 1/5s, 40 milliseconds)

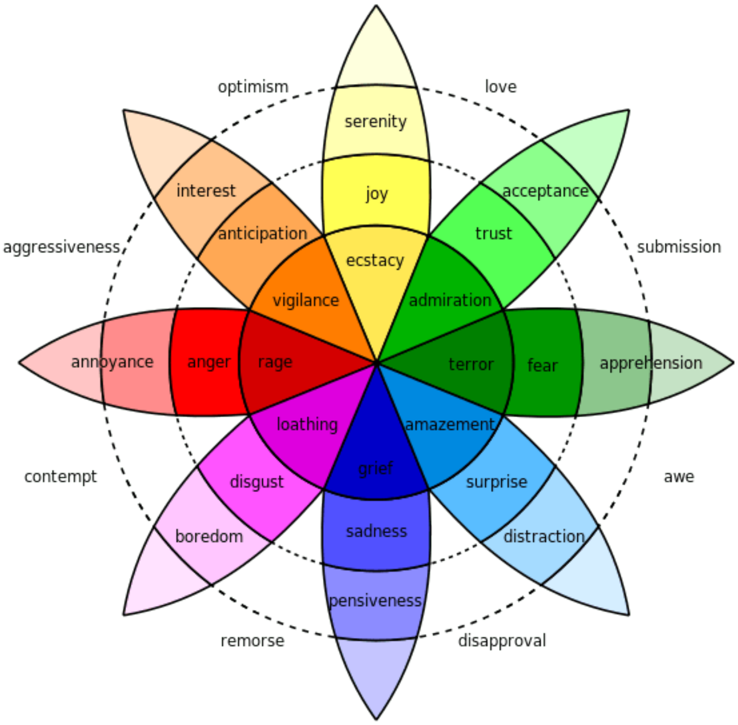


Figure 2.4 Wheel of emotions according to [58] .

and subtle intensity, its recognition is very challenging [64]. Micro-expression recognition methods can be divided into two sections: hand-crafted methods and deep learning-based methods. In the hand-crafted method, [65] applied optical flow method on different subregions of the facial area to recognize subtle movements of facial regions of interest. The directional mean optical flow feature is computed to reduce head movements' effect on the feature vector. This feature vector is fed into the Support Vector Machine (SVM) classifier for micro-expression recognition. Another research is done by [66] to address head movements in micro-expression recognition by presenting an algorithm based on optical flow estimation to perform pixel-level alignment to characterize facial dynamics of micro-expressions. Then, sequences of micro-expressions are divided into a collection of cuboids, and the principal optical flow direction is calculated for each cuboid. These features are passed to the SVM classifier to identify micro-expressions.

The above-mentioned researches have made an important contribution to micro-expression recognition. However, the methods depend on hand-crafted features and mostly extract shallow information, which is not enough for abstract feature description. This process might be difficult for a psychologist with a lack of experience in using such methods. Hence, a method that automatically generates high-level features is desirable. Advances in computer hardware and GPU-based computing have made it possible to use a large dataset to train deep learning models. Recently, few deep learning-based approaches have been explored for micro-expression recognition. Convolutional Neural Network (CNN) used by [67] to detect the facial landmarks' location. Another CNN is then utilised to estimate the optical flow features from facial regions highlighted by the facial landmarks. These features are fed into SVM for recognising and detecting the micro-expression. Transferring Long-term Convolutional Neural Network (TLCNN) has been employed by [68] to extract micro-expression features from each frame in the video clip. These features are then used as input to Long Short Term Memory (LSTM) to learn the micro-expression temporal sequence information. In another study done by [69], dynamic information extracted from optical flow features using 3D flow-based CNNs to learn Spatio-temporal changes by facial movements in the video-based dataset. However, the mentioned studies did not investigate micro-expressions' spontaneous occurrence, which represents an involuntary action toward an emotional stimulus. Therefore, in chapter 5 and 6 of this thesis investigates the events at different time scales. This can be quite significant for mental health because it can model and capture visual features automatically without prior knowledge about the mental condition.

2.1.4 Optical flow

Optical flow's objective is to estimate the movement of interesting features in successive images of a scene. This technique has been used for facial expression recognition from image sequences [70] [71]. Where the movement patterns of the skin are used to extract optical flow motion vectors that discriminate facial emotion changes in a sequence of frames. A typical optical flow algorithm [72] assumes constant brightness, which means that a pixel from the image of an object in the scene does not change in appearance as it moves from frame to frame. In addition to brightness constancy, the Lucas-Kanade algorithm adds temporal persistence assumption, which means that the movement is within a small window (3x3) around the point of interest [73]. However, the disadvantage of using small local windows in Lucas-Kanade is that large motions can move points outside of the local window and become impossible for the algorithm to find the related point movement. To address this limitation, a pyramidal Lucas-Kanade algorithm was proposed by [74] where the motion vectors are calculated at the highest level of an image pyramid. Then, the flow vectors are propagated up to the next lower level of the pyramid. A more accurate optical flow is then recalculated for this new level using the optical flow from the previous level as an initial guess. The resulting flow is propagated one level up. This process is repeated until the flow is calculated at the lowest level.

2.2 Machine learning

Machine learning algorithms are concerned with utilising extracted features to build models that can generalise and recognize the target classes. These algorithms have two variations; supervised and unsupervised learning. In supervised learning, labels and observations are provided; in this case, the classification algorithm aims is to classify the unseen samples. On the other hand, in unsupervised learning, labels are hidden, and the classification algorithm task is to seek a common structure in the observations. In supervised learning methods, discrete or a regression value could be predicted by regressors. The following section discusses feature engineering and machine learning methods that form essential parts of the computational methods for diagnosing mental disorders.

2.2.1 Cross validation

To assess the machine learning algorithm's ability to generalise well to the newly introduced data and reduce the possible bias to the models, cross-validation methods are employed. This is done by dividing the dataset into two parts, one for training and the other is left out of the testing so that training samples are left out of the testing process. Thus, model performance can be assessed when newly unseen data is introduced. There are two methods for cross-validation: leave-one-out and k-fold. In the leave-one-out method, the training set is prepared as *N-1*, which means that training data will consist of all samples at every iteration, but one is left. Then, the algorithm is tested for *N* times to assess its ability to classify the left-out sample. In the k-fold method, the dataset is split randomly into *K* partitions, e.g., for *k* = *10*, each time one partition is kept for testing and the other partitions used for training. This is repeated for *K* times. This process is executed for *N* repetitions, where *N* is the whole number of samples.

2.2.2 Feature fusion

Typically, several visual features can be extracted for affective sensing applications, such as appearance-based features, transfer learning features, and statistical features from time-series observations. To use these features, the feature fusion techniques are employed to concatenate features to enhance machine learning algorithms' outcome. There are two fundamental ways to implement feature fusion: feature level fusion and decision level fusion. In feature level fusion, several features are combined to create a larger vector. While decision level fusion is implemented when the classifier decides the input feature vector label. This technique can be implemented in two approaches: hard decision or soft decision. The Hard decision is applied by majority voting on the predicted labels by the classifier using logical AND or OR. While Soft decision fusion is implemented on the classifier probabilistic outputs or by merging non-probabilistic outputs employing suitable weights. Typically, feature level fusion results in poor performance compared to decision level fusion. This is because feature level fusion increases the feature vectors' dimensionality and adds more complexity to the machine learning task. This might make the classifiers and regressor struggle to learn meaningful representation due to increased data dimensionality.

2.2.3 Feature selection

This approach is an essential task in a machine learning pipeline, where a subset of features is selected from the original high-dimensional feature vector. This will help remove features that might negatively impact the learning process and enhance the machine learning algorithm's performance as the model is trained on a reduced set of features [75]. There are three techniques used for feature selection, namely filter-based approach, wrapper approach and embedded approach [76–78]. Filter based approach selects features independently from the classification or regression model based on a given criterion. However, this task depends on the given task (i.e., classification or regression). For example, feature selection based on filter methods for classification includes t-test, Mann-Whitney U-test and minimum redundancy maximum relevance (mRMR) [79]. For regression tasks, Pearson correlation coefficient, Spearman correlation coefficient algorithms are used [80]. The wrapper approach utilises the machine learning model capability to monitor the performance by adding or removing a feature or subset of features. This process is repeated iteratively until obtaining the highest model performance. Examples of this approach are sequential backward search (SBS) and sequential forward search (SFS). Finally, in the embedded method, features are selected as a part of the classification or regression learning procedures. As an example, feature importance is built-in tree classifiers and therefore can select the best subset of features, Ridge regression, elastic-net and partial least squares regression (PLSR) [81]

2.2.4 Feature extraction approaches

Humans communicate their emotional behaviour and mental state through social signals [82]. Among the wide range of social signals, communicating information through the face and speech is most commonly used for developing automated screening methods for mental disorders. Feature extraction is an essential stage in FER work-flow since the following stages completely depend on it. According to [83], feature extraction algorithms can be divided into two techniques a) geometry-based or b) appearance-based. In the scope of mental disorders diagnosis, time series features are obtained of both (a) and (b). Moreover, visual features are classified into the high or low level; high-level features are directly translated to human sense; on the other hand, low-level features provide image processing descriptors. Software packages can be used as feature extraction methods such as OpenFace [84] and Opensmile[85].

2.2.5 Preprocessing

After collecting raw data, it should be pre-processed by removing noise or reducing the high dimensionality to prepare it for the feature extraction stage. The pre-processing stage might include signal enhancement by removing outliers to improve the features extracted from the raw data. Data normalisation or standardisation is another pre-processing technique that adjusts different data values into a unified scale. Moreover, normalisation provides a fair comparison of data samples as it removes variations resulted from differences in recording environments. Several normalisation methods could be used (e.g., Min-Max-Scaling method, Standard-scaling method and Robust-scaling method).

2.2.6 Functionals

This approach uses descriptive statics to summarise low-level features to create features that provide global representation about the visual or audio recordings. These features are calculated based on a consecutive sequence of frames to supply local information. In the upcoming chapters, descriptive statistics functionals such as mean, skewness, kurtosis, maximum, minimum and standard deviation are employed to summarise the extracted features. For example, using maximum functional on the eye gaze feature vector provides information about the maximum eye gaze duration for an individual during video recording. Multiple functionals should provide a global description of the recording as a single function is not enough.

2.2.7 Classification and Regression

Several machine learning classification and regression algorithms have been used for developing a computational framework for the automated screening of mental disorders. For example, support vector machine (SVM) [86], decision trees [87], random forest (RF) [88], extreme learning machine (ELM) [89], and logistic regression [90]. For the task of automated depression and Schizophrenia recognition, SVM has been the most popular algorithm [91]. Also, decision trees [92, 93] and random forest [94, 95] have been used for automated depression recognition. The literature study suggests no particular method for selecting a classification or regression algorithm to develop automated diagnosis methods. This is probably because of the complexity and the subtle duration of expressing human behaviour. For example, there is a large overlap between the classes due to the temporal nature of behavioural patterns in mental disorders. Therefore, Kachele et al. [96] suggested the use of an ensemble of classifiers and regressors for developing automated tools for the recognition of affective states and depression. Since classes are not linearly separable, Kachele et al. [96] believe that using a single classifier result in poor performance as it tries to learn the relationship between contradicting data points or selecting the best set of features that may not represent the given task. The amount of available data is another aspect that affects the choice of classification/regression algorithm. A brief introduction of classification and regression algorithms used in this thesis is provided in the next sections. These include SVM,RF, Adaboost and deep learning.

2.2.8 Support vector machine

The Support Vector Machine (SVM) was introduced by Cortes et al. [97] for binary classification tasks. Drucker et al. [86] introduced Support vector regression (SVR) for regression tasks. SVM aims to find the optimal hyper-plane that finds a maximal margin between two or more classes in terms of classification. As shown in Figure 2.5 (a), the algorithm finds several hyper-planes, but only the optimal plane differentiates the two classes. SVM can be used for nonlinear classification/regression tasks by extending the feature matrix into a hyperspace using a suitable kernel such that linear separability exists in that hyperspace before using

SVM. The task of selecting a suitable kernel requires cross-validation for tuning parameters of the various kernel functions. Commonly used kernel functions include polynomial kernel, Gaussian kernel and linear kernel [98].

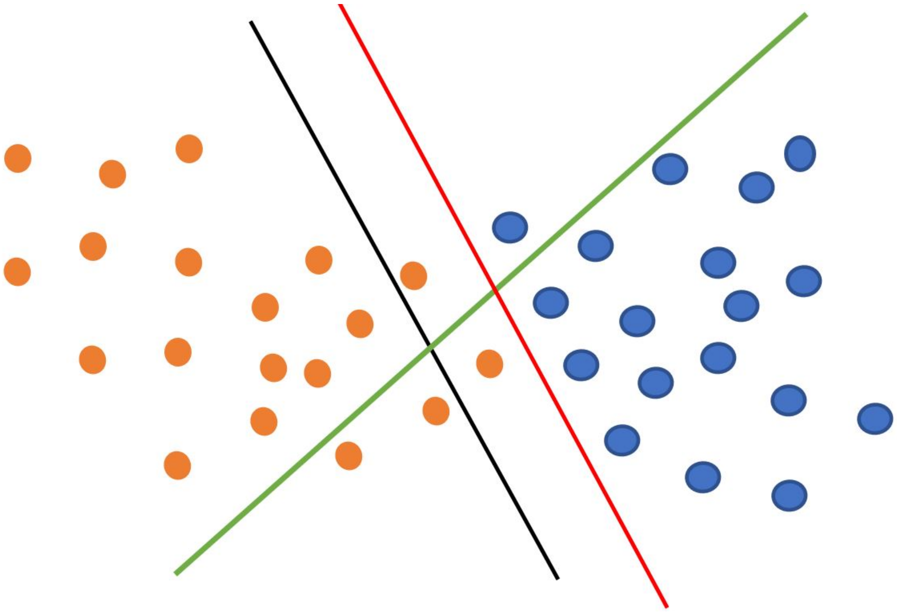


Figure 2.5 Example of SVM classification approache.

2.2.9 Random forest

Random forest (RF) creates an ensemble of decision trees randomly based on the ’bagging’ method, which increases the model performance by combining all learning models, as shown in Figure 2.6. This is done by creating a series of decision trees and combine the output from these trees. RF creates a group of decision trees and combines their output to deliver a more accurate prediction.

2.2.10 Adaboost

The Adaboost algorithm was first introduced by [99]. This classifier has an iterative procedure that builds a strong classifier by combining many weak learners and inaccurate rules. Adaboost uses an unweighted training sample to build a classifier, where if the data sample is misclassified, the weight of that training sample is boosted. Next, a second classifier

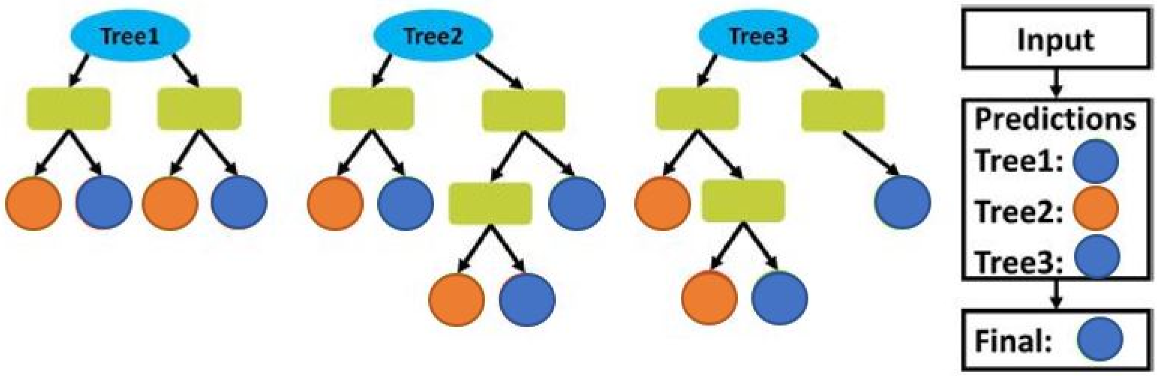


Figure 2.6 An example of Random Forest (RF).

is built using the new weights, and the process is repeated. A score is given for each classifier, and the output classifier is given as a linear combination of the classifiers from each iterate.

2.2.11 Deep learning

Recently, deep learning algorithms have gained popularity in the field of computer vision application due to their success in solving different computer vision problems such as object classification and detection with state-of-the-art performance [100] [39] [101]. Following this success, several researchers used deep learning for facial expression recognition. Deep learning algorithms do not have feature extraction and model learning steps. Features are learned automatically in a hierarchical manner by multiple layers of neural networks. Convolutional Neural Networks (CNNs) are the most used variation of Artificial Neural Network (ANN). These networks have shared weight architecture and local connectivity between neurons. The first study that employed CNN for facial expression recognition was [102], where 6 CNN layers with two types of architectures were used. The first architecture with fixed filter size and the 2nd with different filter sizes. These multiple filter sizes enable the network to extract features at multiple scales. However, this study used a sigmoidal activation function which is susceptible to vanishing gradient problem in deeper architectures. Moreover, shape and temporal information were not utilised. A new 3D CNN architecture was introduced by [103] to learn facial expressions in a video sequence. In this architecture, certain dynamic parts of the face are learned and then they were used for facial expression classification. This approach introduced the concept of detecting useful facial parts for FER tasks and using a sequence of contiguous video frames to learn temporal information. However, facial shape information was not used and the time to learn the temporal information was limited to *n* consecutive frames. Hence, temporal information longer than *n* is ignored, which may discard useful information for certain applications. To overcome the problem of a small amount of data, [104] used two deep CNN network for facial expression recognition. The first network is trained on image sequences, and the second network is trained on the temporal facial landmark points. Later, these two networks are combined to find the final decision for the facial image class. This approach presented a new way of modelling temporal shape and appearance; however, they modelled separately in the network. For this reason, the network might not be able to learn the optimal combination of shape and appearance. Also, this study has the limitation of using a limited number of image sequences as input. For FAUs detection, fewer CNN based approaches have been investigated. A deep CNN with 3 convolutional layers was implemented by [105], 1 downsampling layer and a fully connected layer to detect the presence and intensity of FAUs. Results from this study were comparable to other participants of the FERA-2015 challenge [47]. These CNN based approaches [102], [103] and [105] provided separate modelling of key facial features i.e. shape, appearance and dynamics. While [104] utilised all these features but learned them separately and used a fixed frame length to learn temporal information. This might cause losing important information that happens at different time scales.

2.2.12 CNN architecture layers

The architecture of CNN consists of a sequence of stacked layers, followed by fully connected layers. However, recent architectures in [101] [106] have introduced new ideas that consist of multiple layer architectures. Generally, the input to CNN is an image in the form of (height, width, channels). The spatial dimension is represented in this form of input together with feature maps of the input.

2.3 CNN architecture elements

- **Feed-Forward Neural Networks**:

For a classification or regression problem with input x and target y, the aim of Feed-Forward Neural Network (FFNNs) is to map input x to the target y, so that $y = (f^{*}x)$. A feed-forward network defines a mapping $y = f(x;\theta)$, where the best function approximation is obtained by learning the value of the parameters θ from the available observations of x and y [107]. FFNNs are typically represented by a series of many different functions $f(1), f(2), f(3), ...f(m)$. An acyclic computational graph is used to describe the structure of FFNN model, where information moves from input x in one direction passing a network of nodes implementing mapping of f to the output y.

- **Fully connected layer**:

In this layer, the input vector is multiplied with a weight matrix w, and a bias b is added. This operation is usually followed by the application of a non-linear activation function. During the network training process, the values of w and b are learned from data. The network output layer is defined as :

$$X^{l} = ReLu(Wx^{l-1} + b) \tag{2.3.1}$$

where *ReLu* is the Rectified Linear Unit activation function which discards the negative elements of the input [108],W is the weight matrix and b is the bias. The *ReLu* is defined as :

$$ReLu(x) = max(0, x) \tag{2.3.2}$$

- **Convolution layer**:

This layer extracts features from local patches of the sampled data such as image, video, or time series. Each convolution layer contains a set of filters whose parameters need to be learned. This is done by convolving filters with the input data compute activation map describing specific features or patterns in the data. The convolution is usually followed by the application of bias and non-linear activation. The convolution layer filter size, also known as the receptive field, defines the scale of feature extraction and is usually has a fixed size that is smaller with respect to the input. The output of *j*-th filter at 1D convolutional (Conv1D) layer *l* is obtained as:

$$X_j^l = ReLu(\sum_{i \in D} x^{l-1}{}_i * w^l{}_{ij} + b_j^l) \tag{2.3.3}$$

Where D is the number of channels in the input, $w_i^l j$ is the weight for i-th input dimension and b_j^l is the corresponding bias. Convolution filter kernels are designed to encode the types of features being extracted. By stacking several convolution layers, the network can gradually learn feature representation at various time scales. To support this, convolution layers are usually alternated with pooling layers. Using pooling layers downsamples the data and reduces the extracted feature subsets, increasing the next convolution receptive field [109]. Hence, pooling operations downsamples input time series and reduce the extracted subsets. As an example, given a 1D input T d $\in \mathrm{R}^T$ s max pooling operation for 1D with stride *s*, pool size$p$ and zero-padding to length T can be described as follows:

$$y = maxpool(d) \in \mathbb{R}^{\frac{\mathbb{T}}{}} \tag{2.3.4}$$

- **Regularisation** :

  When a machine learning model becomes highly specialised for training observations instead of learning the generative data model, regularisation is applied to reduce overfitting. Gaussian(L2) and Laplacian (L1) are common parameters used to regularise a neural network [110]. Besides these, dropout [111] is another regularisation mechanism applied in deep neural networks, in which a random number of neurons are ignored during training. This will remove their contribution to information flow and weight updates. Dropout reduces the neuron co-adaptation and enforces them to learn a generative model instead of specialisation to specific values, resulting in a better model generalisation.

- **Pooling layer**: The pooling layer aims to down-sample input representation feature maps. This will reduce the computational cost and control overfitting by providing an abstracted form of representation. Pooling operation is applied across the whole input by a sliding window on a single depth slice at a time. In CNN, pooling operations can calculate Max or Average of a given input. Maxpooling is the most commonly used pooling method. This method down-samples the input image by a factor of k_x, k_y along each direction and returns the maximum value only inside the sub-region [112].

- **Rectified linear unit layer**:

Rectified Linear Unit (ReLU) is a widely used activation function in neural networks, especially in CNN. This layer adds nonlinearity to the neurons where it returns 0 if the input is negative and returns the same value for positive input. It can be written as :

$$f(x) = max(0, x) \tag{2.3.5}$$

Activation functions such as tanh and sigmoid are not generally used because they return values between 0 and 1. In this case, when the previous layer's output is near 0, the neuron is ignored, and when the output is close to 1, the neuron saturates. This will make both cases useless. These two functions are slower and more expensive to compute than ReLu as it includes exponential values. ReLu layer is proved to be more effective for using a threshold[100]. To prevent all negative values to be set to 0, the Leak Relu is introduced by He et al.[113] in which a small negative slope is introduced. They have shown that this method can be adjusted to give the best performance.

- **Classification layer**:

The classification layer commonly uses the SoftMax function to measure the fully connected layer's performance using posterior class probability. The SoftMax function is shown in equation 2.4.6, where the output of CNN is formed as a vector of scores z with the total sum set to 1. Then, the cross-entropy loss is found using Equation 2.4.7.

$$S_n(Z) = \frac{e^{z_n}}{\sum_{i=1}^{N} e^{z_i}} \tag{2.3.6}$$

$$L = -S_{yi} + log\left(\sum_{n=1}^{N} e^{S}{}_{n}\right) \tag{2.3.7}$$

2.3.1 Evaluation metrics

To evaluate the performance of machine learning models developed in this thesis, the following commonly used performance metrics were employed:

- Accuracy: refers to the number of correctly classified samples. This metric is used in the majority of studies and based on the following confusion matrix :

  $$\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$$

  TP represents the true positives number, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. Accuracy is calculated according to equation (2.4.8)

  $$\frac{TP+TN}{TP+TN+FP+FN} \tag{2.3.8}$$

- Recall: is the measure pf the number of samples classified as true positive in proportion to all objects [114]. It is beneficial for reporting the percentage of the positive cases that were detected. Hence, a high recall score indicates that a few samples were missed [115]. To find the Recall, the following equation is used :

  $$recall = \frac{TP}{TP+FN} \tag{2.3.9}$$

- Precision:

  is another metric used for measuring performance. It measures the number of objects classified as true positive in proportion to all positive objects [114]. In other words, precision shows what percentage of the positive predictions were correct. This measurement indicates the algorithm's ability to make correct true predictions without falsely admitting objects [115]. To find the Precision, the following equation is used :

$$precision = \frac{TP}{TP+FP} \tag{2.3.10}$$

The F1 score metric is also used to report the results. It is an overall measure of a model's accuracy that combines precision and recall. F1 is given by:

$$F1 = 2\frac{precision.recall}{precision+recall} \tag{2.3.11}$$

- Unweighted Average Recall (UAR): is the mean of class-wise recall scores and commonly used as a performance measure, instead of accuracy, which can be misleading in the case of class-imbalance.

- Regression metrics: The depression score for a video recording is obtained as an average of the video frame sequences' predicted scores. The overall performance can be estimated using two commonly used evaluation metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE):

  The MAE and RMSE are defined by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - x_i\right|$$

(2.3.12)

$$RMSE = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(x_i - x_i)^2}$$

(2.3.13)

2.4 Psychomotor motor changes in mental disorders

Psychomotor activities are skills that require coordination between the brain and the body to function effectively. These activities cover multiple domains, including speech and body movements and how activities are affected by mental state and emotions [116]. Mental disorders negatively affect psychomotor activities, suggesting that analysing psychomotor activities can help diagnose mental disorders. For example, depression and bipolar can be identified by monitoring psychomotor symptoms according to the DSM-5 manual [15]. In the case of mental disorders, there are two types of psychomotor symptoms. These include (a) Psychomotor retardation and (b) Psychomotor agitation.

2.4.1 Psychomotor retardation

Psychomotor retardation leads to the slowing of psychomotor activities. From a motor view, it shows impaired speech, idle body movements and fatigue. From a cognitive view, psychomotor retardation can cause impaired thinking and, most importantly, blunted show of affect and emotions.

2.4.2 Psychomotor agitation

Psychomotor agitation is a symptom related to a wide range of mental disorders. Subjects with psychomotor agitation show tension with increased motor movement. Examples of motor agitation are undertaking repeated movements, restlessness, and rapid talking—cognitive agitation manifests as high arousal and negative valence. Psychomotor agitation often found in major depressive disorder or the manic phase in bipolar disorder [117].

2.5 The diagnosis of mental disorders using visual features

The face demonstrates a multitude of information about an individual gender, age and what they are feeling. The effectiveness of visual features study for automatic diagnosis of mental disorder is confirmed in research literature from the field of psychology [30, 31]. In fact, several mental disorders are manifested as altered facial activity. Characteristic features that occur due to psychomotor changes in subjects with mental disorders can be utilised for automated diagnosis of these disorders. This has inspired researchers to study visual features for the automated diagnosis of autism [35, 36], depression [19, 34, 118, 119], schizophrenia [32, 33], and BD [120, 121]. Visual features can be divided into two types, (a) appearance features and (b) emotional features. According to the literature, facial movement analysis can be employed to capture psychomotor activities, whereas emotion expression analysis can be used to experiment with the hypothesis that subjects with mental disorders express their emotions differently.

2.5.1 Appearance features

Studies into possible markers of mental disorders have explored appearance features as a potential physiological indicator of disease advancement, severity, or treatment effectiveness [122]. Facial muscle movement manifests as changes in skin texture with the occurrence of facial wrinkles and changes in the eye region. This has inspired the researchers to study facial muscle movement by extracting visual features concentrating on the face area's texture. To use visual features to study mental disorders, video pre-processing on a frame-by-frame basis is required. Usually, the first step for tracking visual features is to use a face detection algorithm like the Viola-Jones method [37], to find the face region for further processing. Following this step, a registration process is applied to transform video frames into a set of pre-defined pixels. Then, feature extraction step is applied using several algorithms such as Histogram of Oriented Gradients (HOG)[123], Local Binary Pattern (LBP) [124] and Local Phase Quantisation (LPQ) [125]. Facial landmarks detector is another algorithm used to track contour points of key positions on the face, such as the nose, eyes, mouth, chin and eyebrows. Facial landmarks are commonly used to extract geometric feature descriptors for the face by calculating the distance between facial landmarks. This type of feature extraction requires that facial landmarks be already computed for each video frame. Geometric facial features were utilised by Alghowinem et al. [[19], [126], [127]] who have made important contributions to the automated screening of depression. These studies' findings have made important contributions to the automated screening of depression using head movements, eye gaze, eye blinks, and facial landmarks as features to classify individuals into healthy and depressed. To use facial landmarks features, they compute a comprehensive set of distances between facial landmark points, computing velocity and acceleration from the distance calculations. These are followed by calculating the functional summaries to capture facial muscle activity. Similarly, head motion features were computed using velocity and acceleration contours from head movements in terms of yaw, pitch, and roll, followed by functional summaries. To study eye gaze, images of 45 participants were manually annotated to build a 74 point active appearance model around the eye [128]. Then, distance measures were used to describe eye openings, vertical and horizontal eye gaze. This study indicates that individuals with depression have slow motor activity, smaller facial muscle movement and slower head movement. Moreover, they found that depressed subjects show contact avoidance through head pose and eye gaze features. However, their study's main caveat is to calculate a comprehensive set of distance features from 68 facial landmarks manually. Importantly, [19] employed person-specific active appearance models (AAMs) to detect facial landmarks. As AAMs need to be pre-trained for each individual, their use in clinical settings can be difficult, especially with the expected increase in mental disorder diagnosis requirements. Lastly, the reported results were on their private dataset; therefore, they cannot be verified independently.

Similar to Alghowinem et al. [126], Dibeklioglu et al. [118] used facial dynamics features for depression recognition. In their approach, each facial landmark's movement was computed for the entire recording, resulting in 98 different time series features (i.e., the coordinates of x and y for 49 facial landmarks). Then smoothing algorithm is applied with Principal Component Analysis (PCA) to reduce the dimensionality to 15 time-series features for each video frame. Next, velocity and acceleration contours are computed and divided into segments of increasing and decreasing values to apply functionals for feature summarisation. Dibeklioglu et al. used Minimum Redundancy Maximum Relevance (mRMR) [79] for feature selection. They replicated the method of Alghowinem et al. [126] on their own dataset and reported better results in term of classification accuracy. Suggesting that the performance of Dibeklioglu et al. [118] improved due to using mRMR for feature selection algorithm compared to t-tests used by [126]. Pampouchidou et al [119] tracked muscle movements using Motion History Images (MHI) [129] of facial landmarks. A subset of landmarks representing eyebrows, eyes, mouth and nose tip was used instead of all landmarks. For head movement analysis, they calculated acceleration and velocity contours of specific landmarks 2,4,14,16 which are located on the face's contour. Lipton et al. [130] reported abnormal horizontal pursuit eye movements in depressed persons compared to healthy controls. This was confirmed by [131] finding that rates of saccades eye movement correlated strongly in controls but are decreased or missing in patients with mental disorders. This confirms that abnormality in the patient's motor system is a form of psychomotor retardation. Crawford et al. [132] found the same abnormal eye movements in patients without being under medication. Another study done by [133] reported that compared to healthy controls, patients with depressive disorder showed significantly abnormal eye movement indices. Also, the patient's anxiety and depression symptoms and eye movement indices were correlated. Therefore the pathological meaning of these phenomena deserves further exploration. Additionally, Sweeney et al. [134] identified significant eye motor and cognitive performance disturbances in depressed subjects. Nonverbal behaviour, including eye blinks, eye glances and brow-raising are hypothesised to provide a potential data source for discovering a patient's mental state such as the depression [135]. Depressed individuals were found to show decreased eyebrow movement and avoiding direct eye contact with the interviewer than healthy controls [116]. Moreover, a study done by [136] showed elevated blink rates in depressed individuals, which returns to normal as the condition improves.

For eye blink features, Pampouchidou et al. [119] computed the area of facial landmarks around eyes and applied an experimentally determined approach to detect eye blinks. Finally, functionals were used to summarise velocity and acceleration features. This study found that visual features provide a mean F1 score of 0.58 using a combination of training and development set, 0.70 using development set only and 0.47 on the test set. However, features computed from landmarks located at the edge of the face make Pampouchidou et al. [119] approach highly prone to the facial tracker's failures. Yang et al. [93] performed registration of facial landmarks using the mean shape computed using 51 static landmarks from the training, development and test sets. Next, distance and angle measures are computed for eyebrow, eye and mouth regions. Also, Yang et al. [93] investigated the use of AUs features over the entire video recording. The authors reported that their winning submission for the challenge was based on manually designed features from interview transcripts. Huang et al. [137] computed the cross-correlation and multi-resolution between sequences of facial landmarks, head pose, AUs and eye gaze resulting in F1 score of 0.73 on the development set. However, their best results on the test are from interview transcripts. Williamson et al. [138] used the correlation structure to computed the dynamics of FAUs during speech achieving an F1 score of 0.53 on the development set. Lucas et al. [57] utilised video recordings from the Distress Assessment Interview Corpus (DAIC) [139], and they found a high correlation of eye gaze with frowning and smiling concerning the Patient Health Questionnaire PHQ [23] scores. However, the dataset contains only 6 subjects, which compromises the validity of their results.

2.5.2 Emotional features

As individuals with mental disorders typically show altered affect, this fact can be utilised to analyse the emotional expressivity to identify them from healthy individuals. Using facial expressions analysis is based on Ekman's theory for emotions [140]. Scherer et al. [141] used Computer Expression Recognition Toolbox (CERT) [142] to study anger, disgust, contempt, fear, joy, surprise, sadness, and neutral emotions. These examine how emotional neutrality and emotional variability changes according to the depression severity of patients. Scherer et al. [141] study reported a Pearson correlation value of 0.198 for emotional neutrality and 0.054 for emotional expressivity. This indicates that as depression severity increases, flat affect becomes prevalent. Another study done by [143] shows that depression, anxiety and distress lead to a reduction of smile intensity. Stratou et al. [144] studied the emotions in terms of AUs rather than Ekman's six emotions. They found that emotional expressivity is highly subject to the gender of the individual. They report that men have increased activation of AU4 compared to women. This means that men smile more under depression compared to women. A similar trend has been observed for disgust for both individuals with depression. Vijay et al. [145] used OpenFace toolbox to extract AU intensity features to study facial expressions on several forms of mental disorders, including depression. They found that eye-widening and smaller eye openings showed by individuals with depression. Also, they report that brow lowered feature intensity to be distinguishing between depressed and non-depressed individuals.

2.6 Advantages of an automatic screening tool

Mental condition diagnosis is not a straightforward process due to the overlapping of symptoms between these conditions. Additionally, the screening process can be affected by decision bias and accuracy of patients reporting their symptoms. Besides, the current diagnostic tools used for individuals more likely to develop dementia are expensive and invasive and some tests have the risk of radiation. Therefore, developing non-invasive, automatic and objective diagnostic tools that healthcare providers can use repeatedly is highly required. This tool can speed up diagnosis and provide the right medication for patients who are most likely has mental disorders.

2.6.1 Summary

In this chapter, several studies have been reviewed. These studies' objective was to use visual appearance features to develop automatic diagnosis systems to detect early signs of mental disorders. These studies have used a variety of visual features to develop FER systems for mental disorder diagnosis including facial expressions, facial micro-expressions, and eye features [[19], [126], [127],[119],[146]][141],[118, 145]. This is followed by introducing feature extracting and preprocessing approaches. Also, feature selection and feature fusion were introduced, which have been reported in previous studies to enhance automated diagnosis methods' performance. These studies utilised various machine learning algorithms classification and regression algorithms such as SVM, random forest, CNN, LSTM deep neural network, SVR, etc. Moreover, several validation approaches were used, such as k-fold cross-validation and leave-one-out, to show more generalised results. Several limitations were found in these studies, for example, using person-specific facial landmark tracking [19] or unbalanced dataset [23]. These limitations will be addressed in the following chapters when developing the proposed systems introduced, and the results were compared.

Chapter 3

Eye blink detection

This chapter investigates the feasibility of developing a robust eye blink detection system based on utilising facial landmarks to develop a low-cost eye monitoring system to diagnose mental disorders. As mentioned in chapter 2, section 2.6.1, [19] used AAMs to extract eye blinks features and use them on an automated method for depression diagnosis. This method's limitation is that AAMs needs subject-specific training using extracted images from the subject video interview. This requires to extract images and annotate them manually for each video interview; for example, [19] used 45 images for each subject. This process is time-consuming and challenging in clinical settings, especially with the expected increase in mental disorder cases. Therefore, to investigate the role of eye blink features in mental disorders as accurately as possible, this chapter introduces a novel technique to detect eye blinks based on automatic tracking of facial landmarks. Automatic facial landmarks detectors are trained on an in-the-wild dataset and show outstanding robustness to varying lighting conditions, facial expressions and head orientation. The proposed technique estimates the facial landmark positions and extracts the vertical distance between eyelids for each video frame. Next, the Savitzky-Golay (SG) filter is employed to smooth the obtained signal while keeping the peak information representing eye blinks. Finally, eye blinks are detected as sharp peaks, and a finite state machine is used to check for the validity of the detected blinks based on their duration. The efficiency of the proposed technique outperformed the state-of-the-art methods on four standard datasets. The rest of the chapter is organised as follows: section describes eye blink algorithms in the literature and provides implementation details. Section

3.2 introduces the proposed system pipeline and describes each component. Section 3.3 shows the proposed eye blink algorithm results using four video datasets and compares other studies using the same datasets. The last section 3.4 is the summary and conclusions.

3.1 Introduction

Recently, eye blink detection has been used in various applications such as the interaction between disabled people and computers [147], drowsiness detection [148] and cognitive load [149]. Eye blink defined as a rapid closing and reopening of the eyelids and it typically lasts from 100 to 500 ms [150]. Viola and Jones' algorithm [37] employed on most methods to detect face and eyes. However, the Viola-Jones algorithm handles every video frame separately when detecting faces or eyes within a continuous sequence of images. AS there is temporal information between consecutive frames, this makes bounding detection boxes have various sizes and positions despite stable faces. Region tracking is frequently combined with Viola-Jones to achieve higher detection accuracy regardless of the changes in facial pose [151]. Different techniques for blink detection were proposed, and they can be classified into several categories:

- Frame-based eye blink detection :

  In this type of eye blink detection method, eye images classified into open, close or estimating the degree of eye openness. Eye feature descriptor used by [152] based on the variants of the histogram of oriented gradients. These features were classified using SVM to detect eye openness and achieved about 90.37% accuracy on Zhejiang University (ZJU) dataset. Another approach introduced by [153] to detect eye blinks by extracting feature descriptors from the eye image based using Local Binary Pattern (LBP). These features are fed into SVM and MultiLayer Perceptron (MLP) to interpret the described image features achieving up to 95% accuracy. Due to the size limitation of the available eye blink dataset, [154] introduced a transfer learning strategy for the deep neural network model to improve eye state detection accuracy with a reported accuracy of 97% on ZJU dataset. However, this method detected only eye state (open or close) using static images extracted from videos that limit usability as an eye blink detection method.

- Appearance-based blink detection :

  In this type of eye blink detection, the eye region's textural features used to extract useful information for classification. Contour circle fitting utilised by [155] to test for eye pupil presence for eye blink detection. This approach achieved 96.6%accuracy on a dataset recorded in a laboratory environment. A Weighted Gradient Descriptor (WGD) was introduced in [156] where a new localisation scheme used to validate the eye region returned by cascade models. This approach based on calculating the partial derivatives for each pixel within the localised eye region over time. Weighted vectors obtained in orientations (up and down) and an input waveform is obtained by finding the vertical difference between the y-coordinates of those vectors. After noise filtering, negative and positive peaks in the signal represent the eye's closing and opening, a local maximum and minimum represent eye blinks. The authors in [156] report the best-obtained results for given datasets using different parameters. A new dataset of five people recorded using a 100 fps Basler camera was also introduced in [156]. The reported detection rate on the Basler5 and the ZJU datasets was around 90% and 98.8%, respectively.

- Motion-based eye blink detection

  Rather than depending on appearance features, two or more consecutive frames are needed for frame calculation. A method for using optical flow to analyse the level of angular similarity in orientation between the motion vectors in the face and eye regions described in [157]. This method has been tested on a set of static images rather than video recordings, achieving an accuracy of 96.96% using author made dataset. Drutarovsky also used the Lucas-Kanade tracker, and Fogelton [158] to track the eye region. Around 255 trackers were placed over an eye region; these trackers were divided into 3x3 cells. Next, motion vectors are computed for each cell to obtain the input waveforms for a state machine. If the eyelid moved down and followed by upward movement within 150 ms, a state machine detects eye blink. Additionally, [158] introduced the Eyeblink8 dataset, which is characterised by vivid facial mimics of recorded people. A recall of 73% on ZJU and 85% on Eyeblink8 datasets has been shown in the same research. The authors of [159] used Active Shape Models (ASMs) to obtain 98 facial landmarks. Eye shape is approximated using 8 landmarks for each eye. The average height of eyes to the distance between eyes is used to estimate eye openness. Eye blink is detected if the eye openness degree changes from a threshold larger than 0:12 to a threshold smaller than 0:02. This method cannot deal with more challenging facial expressions found in videos in the wild and uses a fixed threshold for blink detection. Additionally, ASMs need to be trained for each subject individually, which makes it time-consuming to use.

However, these methods are sensitive to image resolution, illumination, and facial movement dynamics. Additionally, the accuracy of the Viola-Jones face detection algorithm is usually reduced in unpredictable scenarios. Therefore, a widely used tracking strategy of facial landmarks is adopted to overcome this problem. Recently, robust real-time facial feature trackers have been proposed that can track a set of interest points on a human face [160]. These trackers have been validated in a battery of experiments that evaluated their precision and robustness to varying illumination, various facial expressions, and head rotation [161].

3.2 Proposed eye blink detection method

The proposed technique for eye blink detection comprises four main steps, as shown in Figure 3.1. These steps are applied to each frame of the input video. Zface [160] is used to localise the eyes and eyelid contours for automatic tracking of facial landmarks. ZFace offers 3D registration from the 2D video without pre-training. The robustness of ZFace for 3D registration and reconstruction from the 2D video has been validated in a series of experiments [162]. A combined 3D Supervised Descent Method (SDM) [160] is employed to define the shape model by a 3D mesh. ZFace registers a dense parameterised shape model to an image such that its landmarks correspond to consistent locations on the face. ZFace is used to track 49 facial landmarks from videos. The eye-opening state is estimated using the vertical distance (d) between the eyelids.

$$d = \sqrt{(P2.x - P1.x)^2 + (P2.y - P1.y)^2} \tag{3.2.1}$$

Where *P1* and *P2* are the eye landmark points obtained from the facial tracker for each video frame. When the eye is open, it is assumed that the obtained distance (d) is mostly fixed, and the distance approaches zero as the eye closes. However, the resulting signal is affected by interference primarily caused by saccadic eye movements and facial expressions. These interference are filtered while the shape of the signal is maintained. Lastly, the filtered signal peaks representing the distance change between eyelids used to detect eye blinks; the higher the peak more likely to suggest a blink.

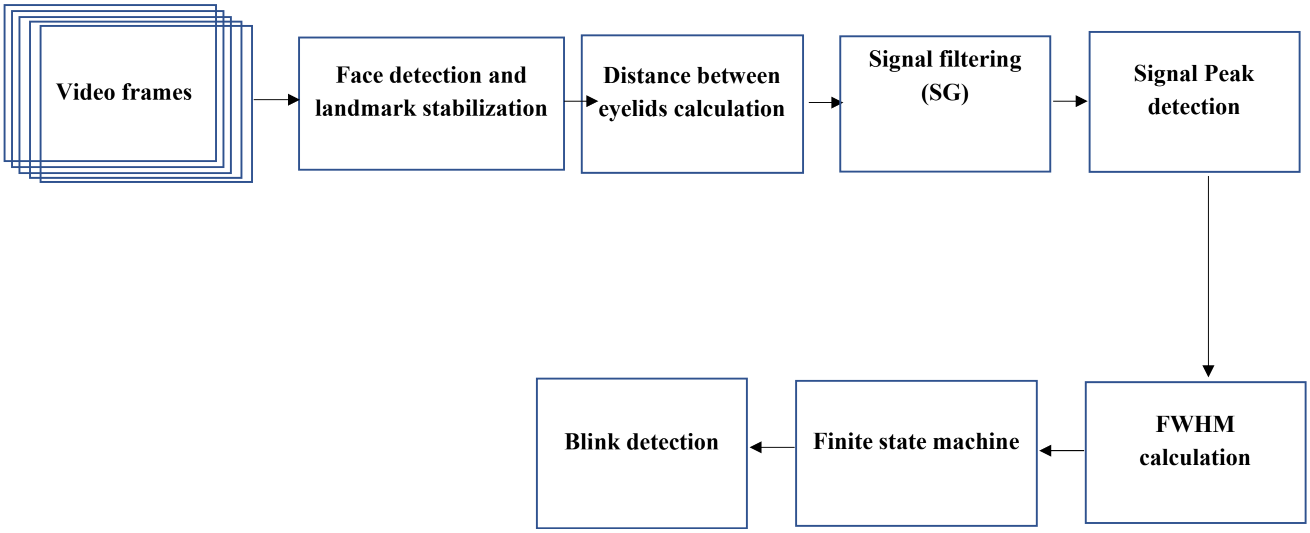


Figure 3.1 Overview of the proposed technique for eye blink detection.

3.2.1 Stabilising landmark points

Factors such as unstable head pose, lighting issues and face obstruction make detection of facial landmarks a difficult task. Moreover, as facial landmarks are detected for each video frame independently, information from one frame is not connected to the next frame's information. This discontinuation of information and unstable landmark points impact applications developed using facial landmark detectors. The landmark stabilisation and tracking method proposed to tackle these limitations, as shown in Figure 3.2, which illustrates the combined landmark detection and optical flow tracking. To implement this technique, the following main steps are used to stabilise the facial landmark detection:

- Face tracking using landmark detector.

- Landmark tracking using Gaussian Pyramid optical flow.

3.2.2 Combining detection and tracking

In the face tracking stage, the facial landmark detector is utilised to track the current frame's landmark location. Next, the Gaussian Pyramid optical flow algorithm is applied to the detected landmark points to predict landmark points in the current frame from the previous frame's locations. These two approaches (landmark tracking and optical flow) are combined in the landmark tracking stage to estimate landmark location in a given video frame. If landmark motion between two frames is small and the tracked landmark's appearance is unchanged, landmark location predicted by optical flow is used. On the other hand, if the landmark locations predicted by the landmark detector and the optical flow tracker are not related, it is assumed that the tracker has lost track and the landmark location predicted by the detector is trusted. However, if this condition dealt with a binary decision to choose one prediction over another, the facial landmark points would be unstable. To solve this, the combination of these two predictions in one method for landmark detection is described below.

Let P(t)= Facial landmark position in the current frame.

P (t-1)= Facial landmark position in the previous frame.

$p_o(t)$= Position of the landmark predicted by optical flow in the current frame.

where α is used to combine p(t) and $p_o(t)$ using the following equation:

$$ps(t) = (1-\alpha)p(t) + \alpha p_o(t) \tag{3.2.2}$$

where α value is $0 \leq \alpha \leq 1$, and the value of α depends on the distance (d) between the location of the point in the current frame (i.e. p(t)) and the location of the point in the previous frame (i.e p(t-1)) as given in the following equation :

$$d = ||p(t) - p(t-1)|| \tag{3.2.3}$$

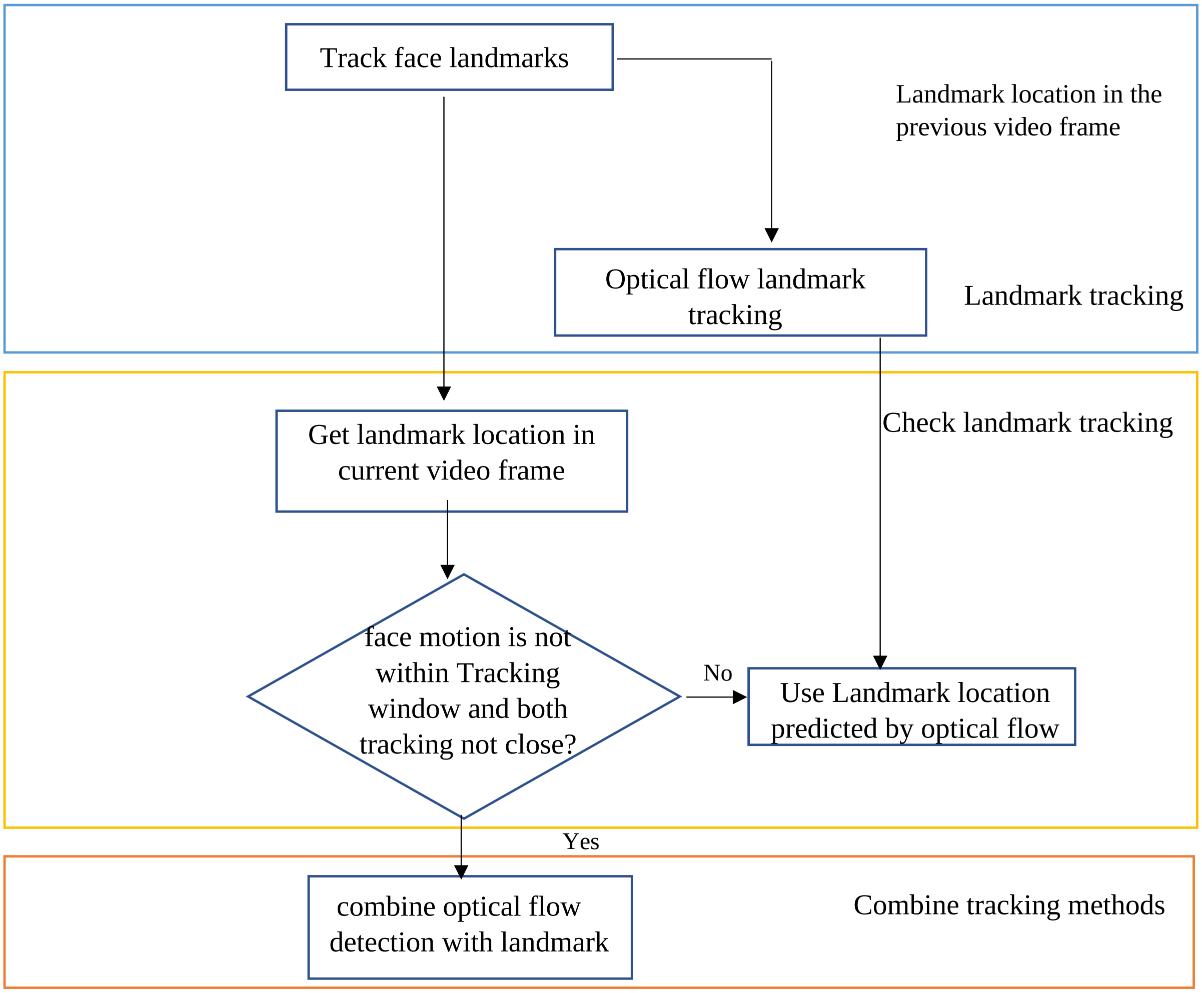

Track face landmarks
Landmark location in the previous video frame
Optical flow landmark tracking
Landmark tracking
Get landmark location in current video frame
Check landmark tracking
face motion is not within Tracking window and both tracking not close?
No
Use Landmark location predicted by optical flow
Yes
combine optical flow detection with landmark
Combine tracking methods

Figure 3.2 Proposed method for landmark stabilisation.

3.2.3 Accuracy of the facial landmark tracker

To evaluate the performance and robustness of facial features tracker, the 300-VW dataset has been used [163]. This dataset contains 50 videos where each frame of a video has the precise annotation of facial landmarks. The vertical distance in equation 3.2.1 used to calculate open, close and other random eye cases then compared with the ground-truth annotation using equation 3.2.4. The average relative localisation error for each face image used to evaluate the accuracy of the facial features tracker, given as:

$$\epsilon = \frac{100}{\xi N} \sum_{i=1}^{N} ||X_i - Xi|| \tag{3.2.4}$$

Where X the ground-truth location of landmark i in the image, i is an estimated landmark location provided by a detector, N is the number of landmarks and ϵ is the Euclidean distance between eye centres in the image. Figure 3.3 and Figure 3.4 shows the robustness of facial tracker on annotated data and detected by Zface [160] tracker before and after stabilising facial landmark.

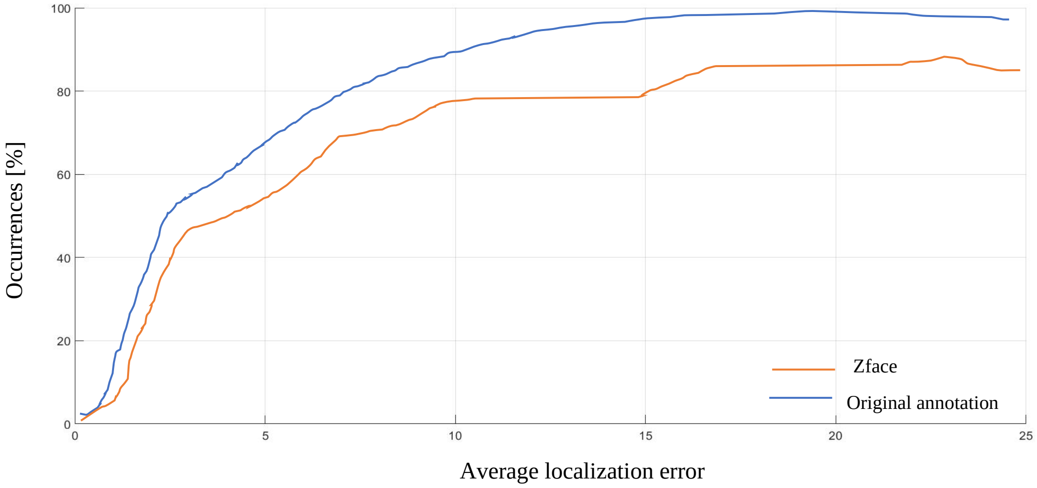


Figure 3.3 Average localisation error of eye landmarks using Zface compared to original annotation.

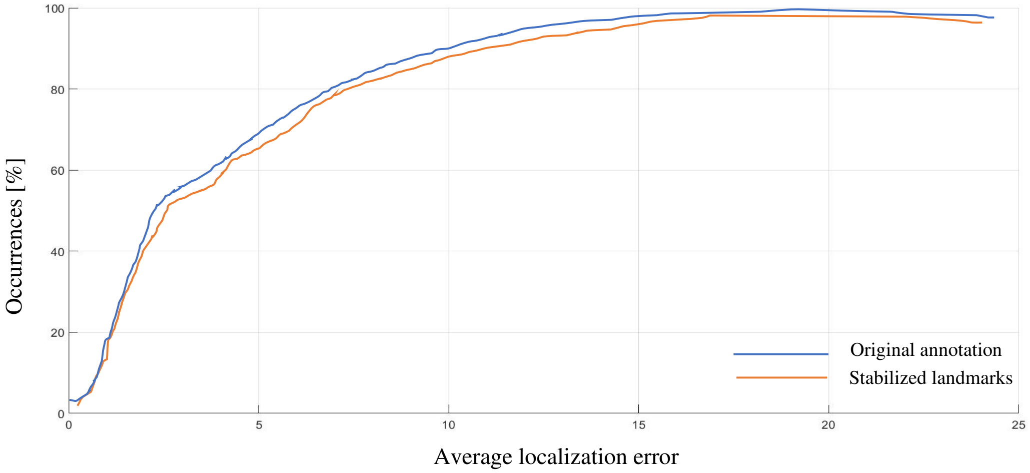


Figure 3.4 Average localisation error of eye landmarks using stabilised landmarks compared to to original annotation.

3.2.4 Pre-Processing of the extracted facial landmarks

In the process of calculating the vertical distance of eyelids, saccadic eye movements, head movements and facial expressions cause unavoidable noise to the signal obtained from equation 3.2.1. To improve signal quality and reduce tracking errors, signal pre-processing is necessary to maintain the signal peaks' shape denoting full eye closure. As shown in Figure 3.5, the obtained signal is noisy and detecting eye blinks; in this case, it may result in false positives being detected as eye blinks. For this purpose, the Savitzky–Golay (SG) filter [164] is utilised for the pre-treatment of the obtained signal as shown in Figure 3.6. The SG filter aims to increase the signal-to-noise ratio without deforming the signal and requires two key parameters: the window size and the polynomial degree. These two parameters are important for reducing the impact of random noise fluctuations and preserving signal information. A long window length causes some loss of valid signals, whereas short window length is ineffective for filtering the signal. Also, choosing a high polynomial degree may produce new unwanted noise, whilst low polynomial degree may lead to signal distortion due to over-smoothing. Therefore, it is important to appropriately select the window length and the polynomial degree to achieve a good trade-off between random noise reduction and valid signal preservation. The polynomial degree is selected in the range of one to three. The window length is automatically adjusted by keeping the polynomial degree constant until an optimal result is found. The highest peak and lowest error are obtained, reducing signal noise while keeping peaks representing eye blinks.

3.2.5 Peak detection

As eye blinks are characterised as peaks in the signal, a median filter is first used for baseline correction to filter out the noise. This noise often caused by certain facial expressions, smiling, closing one's eyes for a time longer than the duration of a blink, and saccadic eye movements. The next step is to detect local peaks in the signal, which are larger than the value of the nearby samples. For 30 fps video, a scanning time window of 500 ms is used to find peaks. Each sample in the signal is detected as a peak if it is the largest value in the scanning window, as shown in Figure 3.7.

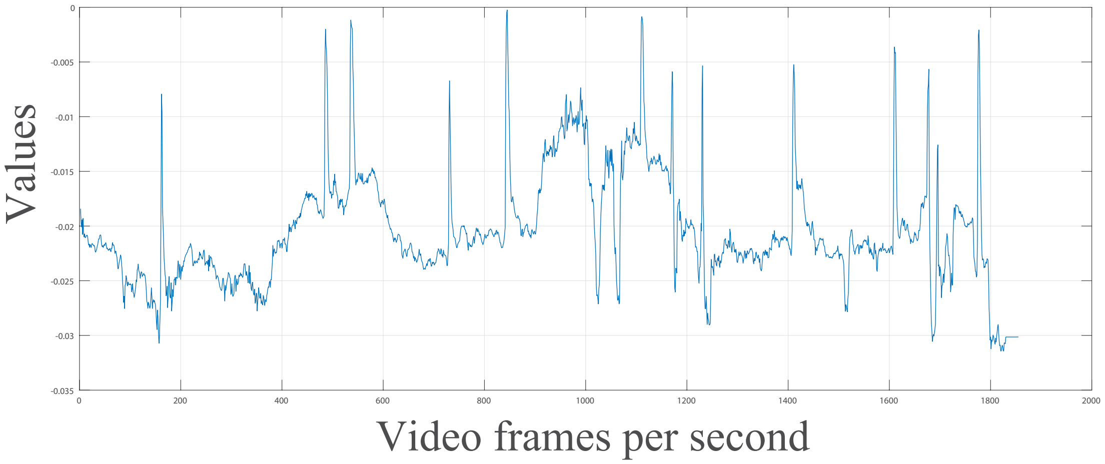


Figure 3.5 Signal obtained from facial landmark tracker in a video tracking session.

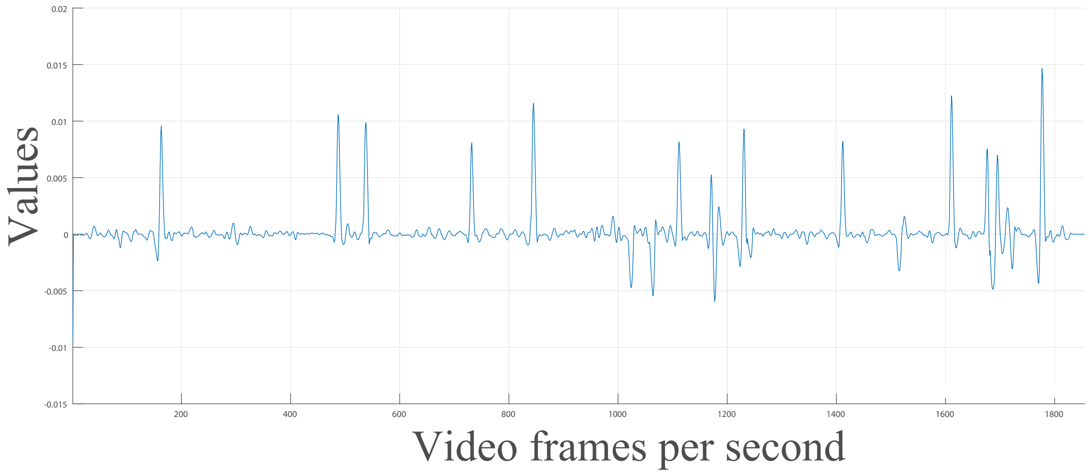


Figure 3.6 SG filter applied to the facial landmark signal

3.2.6 Finite state machine

After the smoothed signal of the eye is obtained; in some cases, false peaks are detected as blinks, as shown in Figure 3.8. a Finite State Machine (FSM) is used to check the detected signal peaks for false and true blink cases. Blink duration is supposed to last for 100-500 ms [150]. Hence, if the eye is closed for more than 500 ms, it is considered that the subject is either drowsy, looking down or blinking voluntarily. FSM is a simple model for keeping track of events triggered by external inputs that consist of states deciding what happens when a particular input is present and which event is subsequently triggered. Then, the calculated peak widths are used by the FSM to differentiate between blink and non-blink peaks according

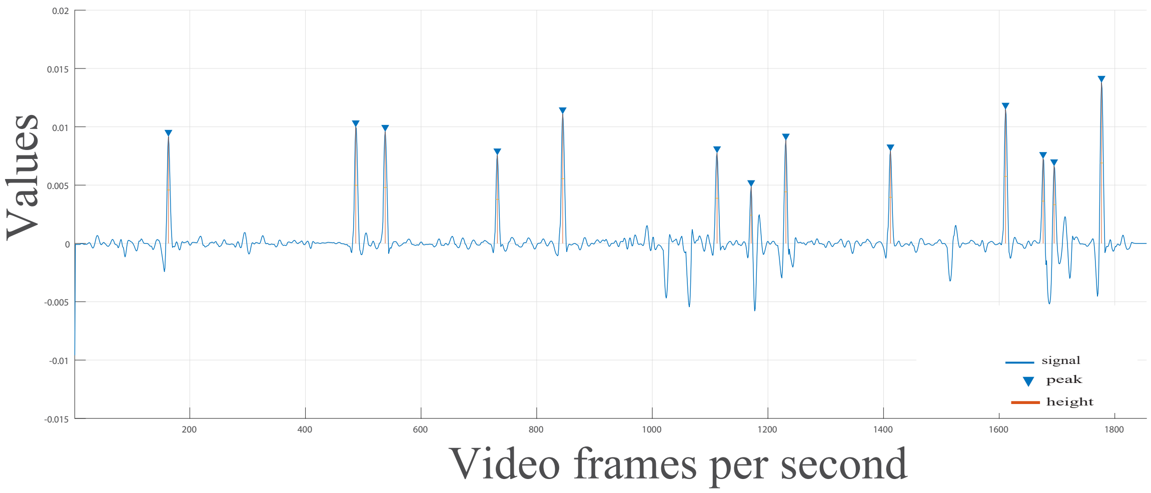


Figure 3.7 Blink peak width calculation using the full width at half maximum (fwhm).

to the following states,

- In-State *S*, if the peak duration lasts for a period of less than 100 ms, it is considered an invalid blink and state are reset to 0.
- In-State *q1*, if the peak period lasts from 100 to 500 ms (16 frames for 30 fps video), it is considered as a valid blink. The state resets to 0, and the blink is validated.
- If the state reaches *q3*, the eye closing period is considered to be more than 500 ms, so blink duration is considered as invalid to state initialised to *S*.

Typically, these states consist of an initial state, input events, output events and a transition function that takes the current state and input event to generate a new output and next state. In this experiment, blink duration is assumed to last from 100 to 500 ms taking into account noise coming from head movement, and facial expressions [150]. The width of the peak represents the time in ms for each detected blink. The Full Width at Half Maximum (FWHM) is used to find the peak width, as shown in Figure 3.9.

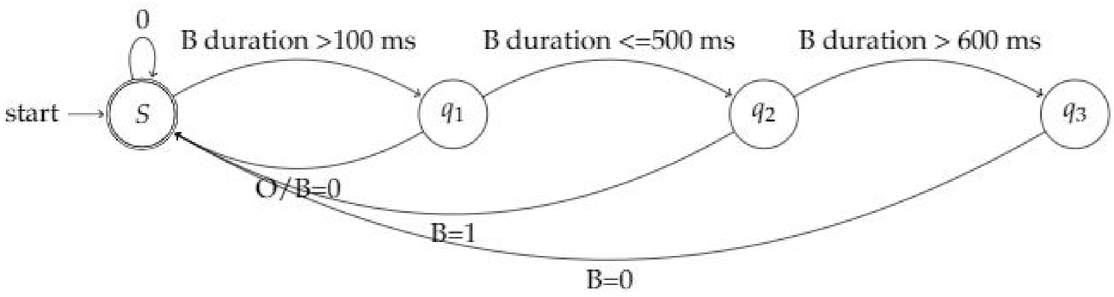


Figure 3.8 Finite state machine for blink duration estimation. B duration: Blink duration.

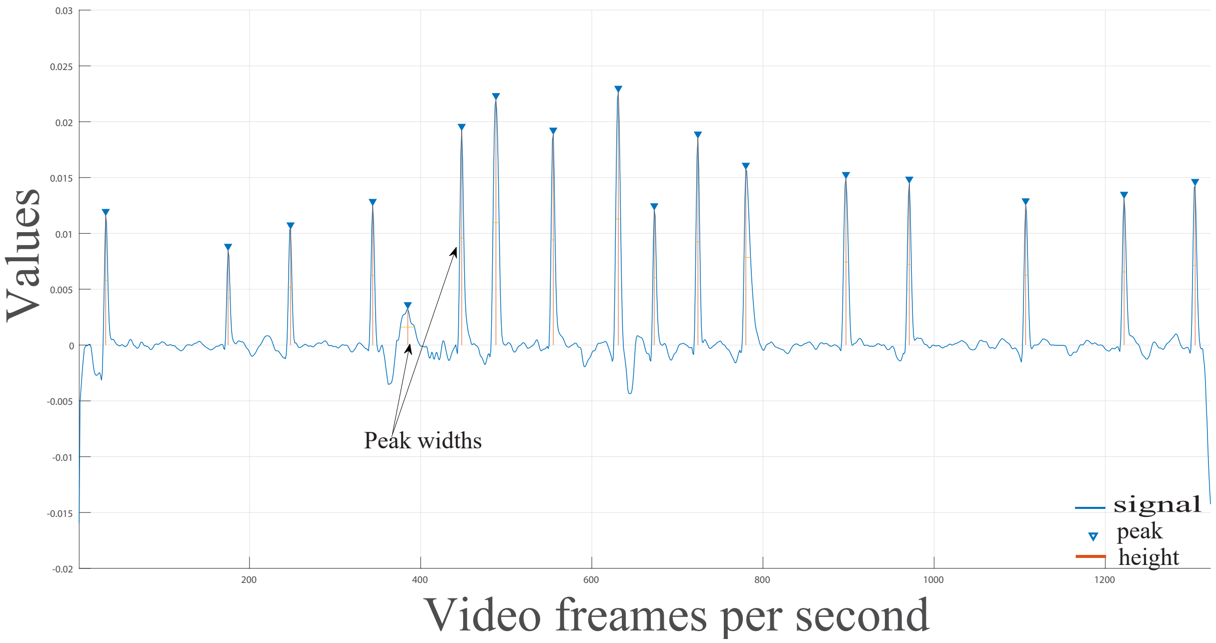


Figure 3.9 Blink peak width calculation using the full width at half maximum (fwhm).

3.2.7 Eyeblink datasets

- ZJU (Zhejiang University): ZJU database [165] consists of 80 videos of 20 individuals. Each individual has 4 clips: frontal view, upward view, with glasses, and without glasses. Each clip is a few seconds and is 30 fps with a resolution of 320 × 240. There is no facial expression and almost no head movements. This dataset has different numbers of ground truth eye blinks reported, as shown in Table 3.1. A ground truth blink is defined by its beginning frame, peak frame, and ending frame.

- Eyeblink8: This dataset is more challenging as it contains facial expressions, head movements, and looking down on a keyboard. This dataset consists of 408 blinks on 70,992 video frames, as annotated by [166] with a video resolution of 640 × 480 captured at 30 fps with an average length from 5000 to 11,000 frames.

Table 3.1 Reported ground truth eye blinks in the ZJU dataset based on different studies.

Reference	**Reported eye blinks**
[158]	264
[166]	261
[156]	258
[151]	255

- Talking face: This dataset consists of one video recording of one subject talking in front of the camera and making different facial expressions. This video clip is captured with 25 fps with a resolution of 720 × 576 and contains 61 annotated blinks [166] [158].

- Researcher's night:

  A new in-the-wild eye blink dataset (Researcher's night 15 and Researcher's night 30) introduced by [166]. This dataset captured at 15 and 30 Frames Per Second (FPS) with resolution 640 ×480. Subjects video were recorded while reading an article on the computer screen. 100 videos were collected from different people with crowded background scene with 1849 eye blinks annotated in total. This dataset is very challenging as participants were acting naturally, some of them warning thick glasses, moving head and showed very challenging multiple blinks.

3.3 Results

3.3.1 Eye blink detection

In this section, the proposed eye blink detection technique's performance is evaluated by comparing detected blinks with ground-truth blinks using the three standard datasets described above. The proposed approach's performance is evaluated following [156], which considers eye blink as detected if the detected eye blink peak is between the start and end frame of the ground truth annotation. Using the mentioned datasets, the proposed method outperforms methods used in [151, 158, 166]. The statistics are listed in Table 3.2. The false negatives counted using the ZJU dataset [165] is because the videos start with a blink and sometimes ends while the subject is blinking. For the Eyeblink8 [158] dataset, the false negatives result from facial expressions, moving hands, narrowed eyes, looking down without blinking, and closing eyes intentionally for a long time.

Table 3.2 Results of the proposed technique and of other existing methods.

Reference	**Dataset**	**Precision**	**Recall**
[166]	ZJU	100%	98.08%
[151]	ZJU	94.4%	91.7%
[158]	ZJU	91%	73.1%
Proposed method	ZJU	100%	98.01%
[166]	Talking face	95%	93.44%
[151]	Talking face	83.3%	91.2%
[158]	Talking face	92.2%	96.7%
Proposed method	Talking face	98.38%	98.38%
[166]	Eyeblink8	94.69%	91.91%
Proposed method	Eyeblink8	96.65%	98.78%
[166]	Researcher’s night 15	92.43%	81 .48%
Proposed method	Researcher’s night 15	98.78%	93.48%
[166]	Researcher’s night 30	81.89%	74 .64%
Proposed method	Researcher’s night 30	97.17 %	90.09%

3.3.2 Eye blink statistics

The main eye blink features are blink rate, duration and amplitude. The blink rate indicates the number of blinks per minute. According to [167], during rest eye blink rate is 17blink/min, increases to 26 during conversation and became very low as 4.5 while reading. Also, [167] showed that the blink rate is affected by the cognitive process more than age, eye colour or other factors. The blink rate in the ZJU [157] is about 50 blinks per minute, because most of the blinks recorded for the ZJU [165] were voluntary and the subjects were informed that they would be recorded for a blink detection study. Blink duration can last from 60 ms up to 700 ms and may reflect the subject’s mental state. Eye blink duration estimation and the total time during which the eyes were in a closed state are shown in Table 3.3.

Table 3.3 Blink statistics: GT (ground-truth blinks), DB (detected blinks), RGT (ground truth blink rate), RD (detected blink rate), and Duration (average blink duration).

Dataset	Video	GT	DB	RGT	RD	Average Duration (ms)
Eyeblink8	1	38	43	4.4	5.1	395
	2	88	89	14.3	14.4	297
	3	65	67	12.7	13.1	222
	4	31	31	10.3	10.3	245
	8	30	34	5.3	6.1	190
	9	41	41	16.2	16.2	224
	10	72	72	14.2	14.2	177
	11	43	44	17.6	18.1	254
Talkingface	1	61	61	24.7	24.7	184
ZJU	261	256	61	48.96	48.03	190

3.4 Conclusion

There has been an increased interest in eye blink detection algorithms for different purposes, such as driver fatigue detection and cognitive load. In this chapter, an overview of eye blink detection methods is introduced, and a novel technique for eye blink detection based on a facial tracker presented. This technique operates using video frames and finding the vertical distance between eyelids, followed by signal filtering using SG, peak detection, FWHM calculation, and FSM to reduce false-negative eye blinks. The proposed methods were experimentally tested and qualitatively evaluated on four standard datasets. The proposed method achieved a higher precision and recall over [166] using all the related datasets. However, false-negative blinks were detected due to unrecoverable noise in the obtained signal; such noise occurred because some of the blinks were voluntary, such blinks often have a relatively long duration, the subject was looking down (e.g. looking at the keyboard), subjects wore thick glasses which reflected light and subjects were too far from the camera. Eye blink duration, frequency and amplitude were also obtained for analysis of eye movement dynamics. In the next chapter, the proposed eye blink detection method will be used to develop automated diagnosis for depression and memory disorders.

Chapter 4

Depression and neurodegenerative cognitive decline assessment using visual features

This chapter investigates visual features' ability to detect the state of two mental disorders: depression and neurodegenerative decline. The developed eye blink detection method in the previous chapter will be used to extract eye blink features to detect person's level of depression using the AVEC2014 depression dataset [168] which according to [19, 21] can be employed to characterise individuals who have depression. The eye blink detection method will be combined with head pose and eye gaze features for neurodegenerative decline diagnosis. Automated diagnosis can provide insights into the mental state of individuals in two ways. The first is to consider the diagnosis as a classification task, where the aim is to detect whether the individual has a mental disorder or not. The second method is a regression task, where the aim is to predict the severity of the mental disorder. The rest of the chapter is organised as follows: Section 4.1 introduces depression assessment, where the current depression evaluation tools, related work, and the proposed automated diagnosis method is presented. In section 4.2, Neurodegenerative disorder diagnosis is introduced followed by diagnosis methods, dataset description, the proposed method and finally, the results were presented and discussed.

4.1 Depression assessment

Clinical depression is related to variety of factors that effect the functionality of the limbic, cortical and subcortical systems. This disorder is a serious medical condition characterised by persistent negativity that disables mental health and daily life abilities of the sufferer, their families and healthcare system. Although those factors may not yet confirmed, they possible to be related to stress and emotional shock [169]. According to Statistical Manual of Mental Disorders (DSM), depression is diagnosed when depressed mental state is present with four or more symptoms listed in Table 4.1 for at least two weeks period [170]. The DSM is widely used to diagnose mental disorder, and its fifth version published by the American Psychiatric Association in 2013. However, it is challenging to confirm a diagnosis using DSM due to its approach in defining the boundaries between the groups and subgroups. As a result, making diagnosis depend on subjective biases in case of valid patient examination can not be done to reach a diagnosis [171]. In a recent report, the world health organisation (WHO) estimated that 350 million people worldwide suffer from depression [172], as shown in Figure 4.1. Moreover, depression is the fourth most significant cause of disability worldwide and is expected to be the leading cause in 2020 [53]. The WHO expects that depression will become the main cause of disease burden within the next 15 years [7]. To assess depression cases, clinical experts need to be involved during the process of diagnosis. Due to the different clinical conditions associated with depression, the diagnosis process is not straightforward. Evidence suggests that subjects could have the same diagnosis despite having different symptoms [173]. These tests describe a range of audio-visual depression symptoms that can diagnose depression by the presence of depressed mental state and obvious less interest combined with at least four of the following symptoms for a period exceeding two weeks: depressed facial expression, hand-squeezing, psychomotor agitation, slowed body movements and neuromotor disturbances; disturbed sleep patterns and loss of energy [174].

Often, these symptoms are not given much attention during screening, diagnosis and follow-up of treatment. However, early diagnosis and re-assessments for following up the result of treatment are frequently limited due to the increasing number of depressed people. This is because depression depends mostly on self-reported symptoms (e.g., the clinician-

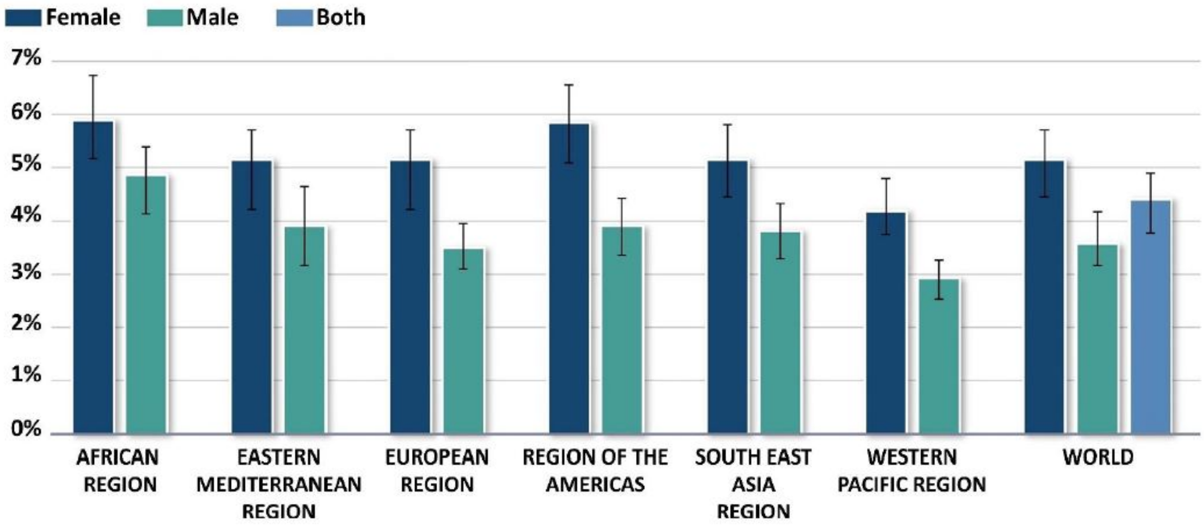


Figure 4.1 Diffusion of depressive disorders (% population), from [176] .

administered Hamilton Rating Scale for Depression [175]). While being useful, these tools lack measures for visual symptoms that are strong indicators of depression. Therefore, automatic depression detection is expected to help objective and timely diagnosis. This will ensure that appropriate clinical care can be given at the right time. Additionally, remote monitoring of depression conditions will be highly appreciated by health care providers. As clinicians assess depression depending on symptoms affecting an individual's daily life, such as the behavioural changes, visual features can be an objective diagnostic tool.

4.2 Depression markers

Identifying depression, biological and psychological markers is the core of the research towards an objective diagnostic tool. Several biological markers related to depression reported, for example, low serotonin levels [24] and genetic abnormalities [25]. However, due to the variety of symptoms, no particular depression biomarker has been recognised. Low serotonin level is considered to be the best biomarker for depression [177]. However, many studies found that healthy subjects from a family with mental illness history [178], or subject with aggressive behaviour [179], might suffer from low serotonin as well. Furthermore, the size of the hippocampal is mapped to depression [180]. Frodl et al. [181] found that 60 patients with depression have small hippocampal compared to healthy individuals. This could point those

Table 4.1 Common depression symptoms.

Depressed and /or Markedly diminished interest or pleasure In combination with four of Psychomotor
Psychomotor retardation or agitation
Diminished ability to think/concentrate or increased indecisiveness
Fatigue or loss of energy
Insomnia or hypersomnia
Significant weight loss or weight gain
Feelings of worthlessness or excessive / inappropriate guilt
Recurrent thoughts of death or recurrent suicidal ideation

gene abnormalities is a marker for depression. Additional biomarkers related to depression are sleep disturbances [182] and saccadic eye motions [131]. However, the current depression diagnosis tools not focusing on non-verbal signals and behavioural markers. Recent studies suggested that these markers can be useful in detecting depression symptoms. These markers include eye blink [19, 183], facial landmark tracking [184, 185] and hand movement [143].

4.2.1 Depression evaluation

The screening tools used in the depression video dataset used in this study are the Hamilton Rating Scale (HAMD) [186] and self-evaluations such as the Beck Depression Index (BDI) [187]. These tools are the gold standard for reporting the severity of depression symptoms. The score reported in each test represents the level of depression severity. These scores are generated differently across these tools according to different weighting methodologies and variety of symptoms covered. These tools share common symptoms including; fatigue, sleep disturbances, depressed state, agitation, guilt, and loss of interest.

Hamilton Rating Scale for Depression

Hamilton Rating Scale for Depression (HAMD) is regarded as the gold standard tool for depression severity assessment. It was introduced for the first time in 1960 [175] and revised for several times in 1966 [188], 1967 [189], 1969 [190], 1980 [191]. This tool consists of 21 questions used to screen adult for depression and assess the severity of their condition by exploring an individual's day to day feelings of guilt, suicide thoughts, anxiety, agitation or retardation and insomnia. The assessment requires administration for 20-30 minutes. Each question has 3-5 outcomes depending on the severity of reported response. HAMD test has five scoring levels; Normal (0-7), Mild (8-13), Moderate (14-18), Severe (19-22), and Very Severe (23).

Beck Depression Index

The Beck Depression Index (BDI) is one of the most widely used self-assessment tools for depression [175]. It contains 21 questions emphasis on cognitive state expressed in depression and the negatives of self-evaluation. Each question is marked as 0, 1, 2, or 3 based on the patient's reported symptom severity over the last week. The highest score is 63 and it has four scales: severe (30), moderate (19-29), mild (10-18), and minimal (0-9). this tool is rated to be convenient to use as the patient can do it without the need for clinicians [192]. However, it depends on the patient's reading ability and self-reported answers which influences its reliability [193].

4.2.2 Related work

It is established that depression appears through nonverbal symptoms in terms of facial expression as well as body language [194]. These characteristics have been researched extensively in psychological studies and have been found that voluntary changes in facial expression, psycho-motor agitation and depressed mental are central to depression [56]. These observations fit well into some of the behavioural diagnostic criteria of depression, as mentioned by the American Psychiatric Association (APA) [15]. Cohn et al. [184], and Mcintyre et al. [195] started early research on depression analysis using facial features from Active Appearance Model (AAM) along with Facial Action Coding System (FACS) to understand how they represented in depressed subjects. These studies concluded that it is feasible to develop a system using facial features for depression analysis. There have been several depression detection challenges to promote further research for depression by introducing Audio/Visual Emotion (AVEC) Challenge series [120, 168, 196, 197]. These challenges have gained interest in their efforts to learn signs of Major depressive disorder (MDD) because this disorder is continually increasing worldwide. In the AVEC2013 and the AVEC2014 depression challenge the data is provided as pre-computed sets of visual features such as HOG, 3D head pose estimation, eye gaze direction estimation, 2D and 3D facial landmark and emotion-based measures. Many participants contribute in the AVEC2013 [197] and the AVE2014 [168] trying to find most useful feature combinations. Methods in [198–200] have shown that patients' visual and vocal features have a close relationship with depression disorder. As depression visual biomarkers are of temporal nature, [19] used frame-based features to extract eye features. In this study, [19] found that depressed subjects show abnormal eye movements compared to healthy subjects. Another study done by [21] found that features from eye region correlated with depressed state and provided the best regression scores. Additionally, [15] found that eye blinks are clinically related to depression bio-markers as it can represent the psychomotor retardation. Saccades, pursuit and rapid eye movements (R.E.M.) are the most eye movements analysed in the literature. A pair of electrodes is placed over the subject's eye to produce an electrooculogram signal for measuring eye movements to study these features. Results from [130] found that depressed participants had abnormal horizontal pursuit eye movements in comparison to healthy controls.

This indicates that the ocular motor systems of patients with affective disorders process eye position error abnormally. Furthermore, according to [134], depressed patients have significant disturbances in neurophysiological processes and eye motor system. Additionally, R.E.M. sleep was analysed in depressed subjects. Kupfer et al. [201] found that R.E.M. latencies are reduced in depressed subjects and suggesting that changes in R.E.M is an objective indicator of depressive disease and correlates inversely with its severity. Eye blink rate features used by [136] for depression diagnosis, the results showed an increasing blink rate in depressed subjects which return to normal levels as their condition improves. Active Appearance Models (AAMs) used by [126] to analyse eye movements in depressed patients. However, AAMs' problem is that it must be trained using image samples from each patient video recording in a subject-specific manner. Using a generic eye AAM model trained using images from different subjects than the current dataset, the model fitting and tracking will not be accurate. To overcome this limitation, [126], used 45 images per subject manually selected from video interviews. This procedure is not well-suited and time-consuming, especially in clinical applications or large numbers of participants. To overcome this limitation in [19, 126] method for blink detection in depressed patients, a pre-trained facial landmark detector is employed to track eye lid movements as presented in chapter 3.

4.2.3 Depression dataset

In the AVEC2014, depression severity challenge [168], the provided dataset consisted of video recordings of individuals with human-computer interaction (HCI) questionnaire task. Two tasks were required to be performed by subjects while the session being recorded through a webcam and a microphone. The first task, which called Northwind, subjects were asked to read aloud a scripted speech. The second task, which called Freeform, was where subjects were required to answer one of several questions: 'What is your favourite dish?'; 'What was your best gift, and why?'; 'Discuss a sad childhood memory'. The dataset is composed of 150 video recordings from 84 individuals. Some of the participants were recorded more than once, 18 participants appear in three recordings, 31 in two recordings and only one recording for 34 participants. The range of ages is between 18 to 63, with a mean age of 31.5 years and standard deviation of 12.3. The duration of recordings ranges from 6 to 248 seconds. Overall, there are 300 recordings given as Northwind/Freeform tasks, divided as 100 recordings in the training set, 50 for the development and 150 recordings in the test set. Before starting the video recording, subjects were asked to complete the Beck Depression Inventory-II (BDI-II) questionnaire. Scores from this test were used to label corresponding video depression severity level, as shown in Table 4.2. Individual's classified by [14] into high- and low-depression severity groups according to the following standard BDI score cut-offs:

Table 4.2 BDI-II cut-off scores and the depression severity level

BDI-II Score	Severity Level
0:13	Minimal depression
14:18	Mild depression
19:28	Moderate depression
30–63	Severe depression

The AVEC2014 depression sub-challenge aimed to predict BDI-II scores from the video recordings. The classification of depressed vs non-depressed investigated by [202] using a 13/14 point cutoff on the BDI-II, resulting in 82% classification accuracy. This cutoff has been used for the proposed automatic depression detection approach.

4.2.4 Selecting the machine learning algorithm

To study eye features in depression, eye blink detection algorithm presented in Chapter 3 is used to extract these features from the video recording, followed by machine learning pipeline shown in Figure 4.2 below. Based on related works [19, 127, 136], three types of features were extracted: eye blink rate per minute, eye blink amplitude and link duration. The proposed method consists of feature extraction, pre-processing and machine learning units. Among several machine learning algorithms, the boosting technique proved to work efficiently on different classification tasks due to its ability to produce a strong classifier by combining a set of weak classifiers that are diverse enough to produce a strong classifier when fused. This technique is done by sequentially selecting weak classifiers from a candidate pool and assign weights to each of them based on an error. In each iteration, the boosting algorithm assigns an importance weight for each classification example; examples with higher weight will get more attention on the next iterations as they are incorrectly on previous iterations. This will tune the weak classification learners towards difficult examples to improve the classifier performance by maximising the training set's margin. AdaBoost algorithm [203] is based on the additive model, which is the linear combination of the base classifiers h$_t(x)$:

$$F_(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{4.2.1}$$

The training samples in the binary classification denoted by (x1, y1), ,(xi , yi), $\cdots$,(xn, yn), where x_i is the i_{th} instance, $y_i \in \{1, -1\}$ is the class label associated with x_i. The algorithm runs for a series of rounds *t=[1,..., T]*. For each sample in the training set, a weak learner produces an output hypothesis whose classification ability is better than random guessing. For each iteration *t*, a coefficient α_t assigned to a weak learner, representing the importance of examples in the dataset for the classification. On each round, higher weights are assigned to the incorrectly classified example. The weights of each correctly classified example are decreased so that the classifier focuses more on those examples [204].

On the other hand, SVM is based on the structural risk minimisation [205]. SVM trains classifier to find an optimal hyperplane that maximises the margin between two data classes, i.e., the distance between the closest sample and the separating hyperplane. Moreover, SVM tries to minimise the wrong classified number of training samples, increasing their generalisation ability. With a SVM, the training set of k training samples represented by x_i and y_i where $x_i \in \mathrm{R}^N$ represents a vector with the attributes of the i_{th} instance, and $y_i \in \{+1, -1\}$ represents the class label for the instance. The SVM aims to find the optimal separating hyperplane (OSH) $w \cdot xi + b{=}0$ between the given classes of data set. The decision function to classify a testing instance x is:

$$f(x) = \ w{\cdot}\, x_i \ + b \tag{4.2.2}$$

To find the OSH, SVM solves the following quadratic optimisation problem [206]:

$$\arg\min \tfrac{1}{2}\|w\|_2 + C\sum_{i=1}^{m} Z_i \quad \text{,subject to } \mathrm{y}_i(wx_i + b) \geq 1 - Z_i \tag{4.2.3}$$

Where w is the vector corresponding to the separating hyperplane, $\frac{1}{\|w\|_2}$ is the margin of the hyperplane, b is a scalar bias term (so the hyperplane is not forced to go through the zero points), z_i are variables classified on the wrong side of the margin of the separating hyperplane called slack variables, and C is a user-defined parameter controlling the trade off between margin and the number of slack variables. The constraints aim to place instances with positive labels at one side of the margin $w{\cdot}\, x_i \ + b \geq 1$, and negative label instances at the other side $w{\cdot}\, x_i \ + b \geq -1$. The optimisation problem's function is a trade-off between maximising the separating margin and minimising the slack variables. When C is large, a higher penalty assigned to slacks, which will produce fewer slacks but a smaller margin.

AdaBoost can automatically find characteristic features for classification from a vast features pool [207]. This could eliminate the need for experts to select informative features from domain knowledge of the given classification task. As SVM maximises the margin while keeping the number of misclassified samples, each sample in the features space is taken into the SVM account when creating the separating hypersurface. However, as the number of features becomes more extensive, the SVM approach for classification will be computationally unmanageable. Due to this issue, advanced knowledge of the most informative features for classification required or features must be selected at runtime. Since this experiment aims to understand the relationship between the eye movements features and depression, manually selecting features may not be suitable. One approach to overcome this limitation is to try all combinations of features and select those with the best classification results. However, this would become computationally expensive.

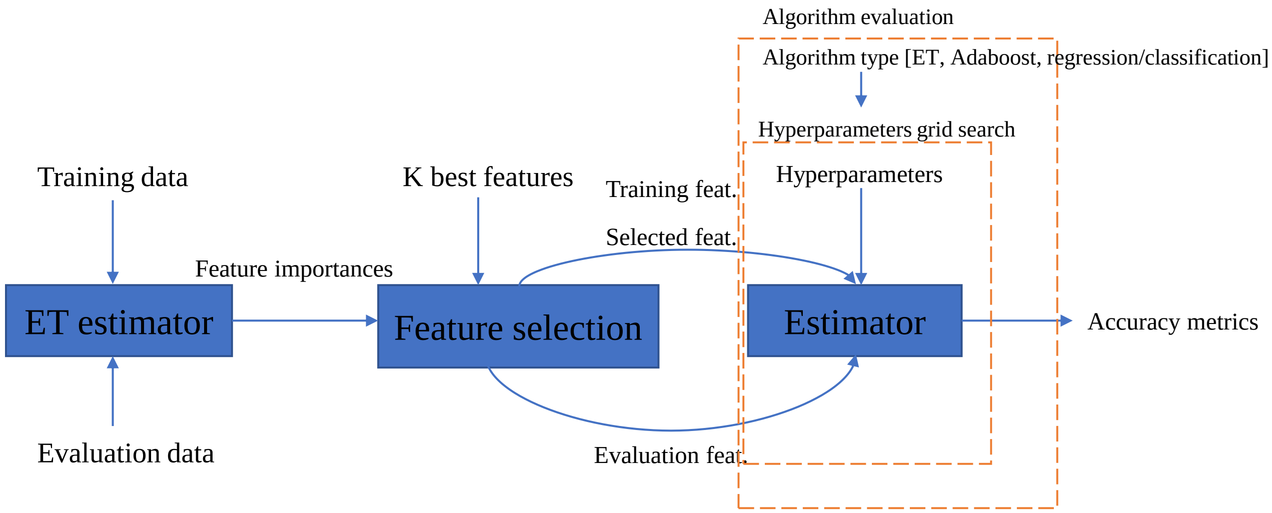


Figure 4.2 Proposed depression detection method.

4.2.5 Feature Preprocessing

Extracted eye features were further processed to obtain several statistical feature description. The extracted functional summaries included average blink duration, closed eye duration rate, closed eye to open eye duration rate and blinking rate for both eyes. Next, the following statistical descriptors where extracted: mean, median, standard deviation, maximum, minimum, skewness and kurtosis. Python version 3.7 implementations were employed for extracting all the statistical descriptors. Before feeding these features into machine learning algorithms, feature normalisation from different scales into a uniformed scale was applied. This gives a unified range to the feature values (e.g. 0,1) to ensure that features have an identical contribution to the classification task and help enhance regression and classification tasks. Normalisation can be done by taking one column simultaneously and finding the maximum and minimum value in each column. Equation 4.1.4 shows min-max normalisation calculation, where X is the feature vector of the i_{th} column, α_{min} is the minimum value of the

i_{th} column, α_{max} is the maximum value of the i_{th} column.

$$X_i = \frac{X - \alpha_{min}}{\alpha_{max} - \alpha_{min}} \tag{4.2.4}$$

Min-Max Scaler method works well for cases when the standard deviation is minimal. However, it is susceptible to outliers, in this case, *RobustScaler* method employed to scale features using statistics that are robust to outliers. This method removes the median and scales the data in the range between the 1st quartile and the 3rd quartile. i.e., in between 25th quantile and 75th quantile range. This range is also called an Interquartile range. The median and the interquartile range are then stored to be used upon future data using the transform method. If outliers are present in the dataset, then the median and the interquartile range provide better results and outperform the sample mean and variance. RobustScaler uses the interquartile range so that it is robust to outliers. Therefore its formula is as follows:

$$X_i = \frac{X_i - median(X)}{Q_3(X) - Q_1(x)} \tag{4.2.5}$$

Where *Xi* is the i_{th} column of the feature vector, *median(X)* is the median of the given observation and the interquantile range is between the 1st quartile *Q1* and the 3rd quartile *Q3*.

4.2.6 Feature selection

To get maximum performance from the machine learning pipeline, features must be explored and examined. Feature selection techniques are employed to find the most important ones and dismiss unnecessary features that affect model performance. Additionally, feature selection may decrease the risk of over-fitting [76]. To determine the importance of the extracted features, the wrapper method is utilised, which uses the classifier's accuracy to evaluate the significance of feature or subset of features. Hence, features that improve model performance will be selected, and others which reduce the performance will be discarded.

4.2.7 Results

Two approaches were implemented for the depression evaluation task. The first approach is eye blink features based regression model to predict the clinical test scores associated with the subject's recording. The second model is a classifier based on eye blink features to identify subjects who suffer from depression from others with minimal or no depression. The three data partitions from the AVEC2014 were utilised to build these models. The model is fitted with samples from the training partition and evaluated twice, against samples from the development partition and recordings from the test partition. This procedure repeated several times until the model finds the best tuning parameters configuration that maximises performance. The baseline results for the visual features model were MAE= 7.57., RMSE= 9.31 for the development set and MAE= 8.86, RMSE= 10.86 for the test set. The proposed model achieved better performance with the MAE= 7.40 and RMSE=9.10 for the developement set, and the MAE= 8.30 and RMSE=10.50 for the test set. These results demonstrate the usefulness of visual features for depression evaluation task. In terms of binary classification method healthy vs depressed, the model accuracy was 93% and 90% for the development and test sets, respectively. However, there were no baseline classification scores reported. Table 4.3 below shows the results for regression and classification using eye blink features. Table 4.4 compares the proposed method with results from the literature.

Table 4.3 The AVEC2014 classification and regression results.

Classifier	Train set	Test set	MAE	RMSE	Accuracy	F1 score	Precision	Recall
SVM	Train: NW	Test: NW	10.5	11.67	60%	0.61	0.59	0.61%
Adaboost	Train: NW	Test: NW	11.2	10.94	75%	0.74	0.74	0.76
SVM	Train: FF	Test: FF	10.8	11.30	64%	0.64	0.63	0.63%
Adaboost	Train: FF	Test: FF	9.8	10.65	88%	0.87	0.86	0.85
SVM	Train: FF+NW	Test: FF+NW	9.4	10.9	61.22%	0.61	0.62	0.61%
Adaboost	Train: FF+NW	Test: FF+NW	8.30	10.50	93%	0.93	0.93	0.92

Table 4.4 Depression evaluation with the AVEC2014 corpus comparison to the literature.

Author	MAE	RMSE	Accuracy	F1 score	Precision	Recall
Baseline [168] video	Dev7.57 Test 8.86	Dev 9.31 Test 10.86	N/A	N/A	N/A	N/A
Baseline [168] audio	Dev 8.93 Test 10.03	Dev 11.52 Test 12.56	N/A	N/A	N/A	N/A
Zhu et al.[208]	7.47	9.55	N/A	N/A	N/A	N/A
kaya et al.[209]	10.26	10.2	N/A	N/A	N/A	N/A
Jan et al. [199]	8.44	10.50	N/A	N/A	N/A	N/A
Pampouchidou [92]	N/A	N/A	Dev 66.0 %	0.72 (weighted)	0.94 (weighted)	0.59 (weighted)
Proposed method	Dev 7.4 Test 8.30	Dev 9.10 Test 10.50	Dev 92% Test 93%	Dev 0.92 Test 0.90	Dev 0.93 Test 0.90	Dev 0.92 Test 0.90

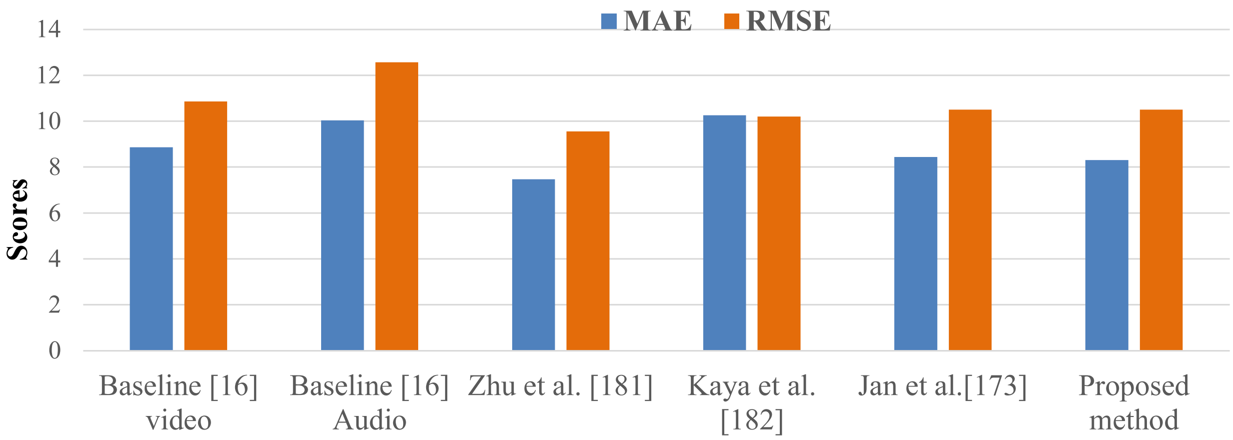


Figure 4.3 The AVEC2014 MAE and RMSE results comparison with models developed using eye blink features.

4.2.8 Discussion

The experiments done in this chapter supports the hypothesis of using eye blink features to diagnose depression and estimate its severity. This is supported by [19, 21, 210] studies, which found that eye blink can be used to identify individuals who have depression. Developing such a system may assist clinicians in their diagnosis and monitoring of clinical depression to provide a treatment which is proven to be effective in many cases [211]. The method described in this research is non-invasive and can be used in a broad range of clinical settings. The SVM and Adaboost algorithms were employed to build this model using the built-in feature importance to select the most informative features. The model accuracies were 93% and 64% for the test set using Adaboost and SVM, respectively. Table 4.3 and 4.4 show the classification and regression models' results using eye blink features extracted from the development, training and test datasets for freeform and northwind tasks combined features. The results show that Adaboost has a better performance compared to SVM. This may be related to the characteristics of the dataset, and the features are not linearly separable. In this case, SVM needs a kernel to project data into a high dimensional feature space where classes are linearly separable. SVM kernel selection can be difficult at a higher dimension because a radial kernel can separate some features, but a polynomial for other features. On the other hand, boosting method adapts to data better than choosing a generic radial basis function (RBF) kernel to project it into infinite dimensions [212], and it has a fast convergence rate, i.e., Adaboost reaches the minimum error very quickly [207]. The video model's baseline results were MAE=7.57, RMSE=9.31 for the development set and MAE=8.86, RMSE=10.86 for the test set. The proposed model outperformed these results on the test set, the MAE=8.30, RMSE=10.50. Additionally, the proposed model achieved better results than the baseline audio model.

Comparing the AVEC2014 results with other studies, the developed model performed better than most previous studies (Table 4.3). In comparison to Pampouchidou et al. [92] study who reported results on the development set only using binary classification, the accuracy, F1 score, precision and recall were 66%, 0.72, 0.94 and 0.59 respectively. While the developed model obtained better results for the same reported metrics 92%, 0.92, 0.93 and 0.92 for the development set, and 93%, 0.90, 0.90 and 0.90 for the test set, with no large difference in the results of the two sets, which means that the proposed system is stable and more likely will be generalised to other datasets. Although in [92], the best scores were obtained with a complex approach that combined both audio and video features, the proposed model achieved better performances using a less complex model. The studies were done by alghowinem el al. [127][19] have shown that eye blink features can help with depression detection. Results presented in these studies were reported using different dataset than the AVEC14 datasets. Hence comparison with their results was not possible. However, individual dependant AAMs have been used in alghowinem el al. studies to track facial landmarks. Since AAMs must be pre-trained for individuals, they are not usable for clinical settings or many participants. In contrast, the proposed eye blink detection in chapter 3 used a fully automatic and person independent method to track facial landmarks. The results of Zhu et al. [208] reported MAE = 7.47 and RMSE =9.55 on the test set using complex deep learning model to model facial dynamics. The proposed model performed better compared to [209] despite the usage of large audio-video feature set by [209] they reported MAE=10.26 and RMSE=10.2 for the test set only.

Results on the development set were not reported by these two studies[[209],[208]]. Therefore, it is difficult to conclude if their proposed model's performance were stable on both sets. Other studies done by Jan et al. [199] used Motion History Histogram (MHH) video features and audio signals to represent characteristics of facial and vocal expression of subjects under depression. Their best results were MAE= 7.36 and RMSE= 9.49 for the development set and MAE= 8.44, RMSE= 10.50 for the test set. Although their best results were achieved with a complex model that combined audio and video features, the proposed model provided better performance using a less complex approach.

Overall, the proposed model reported results to show that eye blink features are useful for predicting the depression severity score. Previous studies show abnormal eye blink features in depressed patients [136]. Additionally, overall longer blink duration has been noticed for depressed subjects. These findings might be a sign of fatigue, which is a marker of depression. This study has a limitation that the AVEC2014 dataset does not contain ground truth annotation of eye blink rate. Despite this limitation, this study demonstrates the feasibility of using eye blink features for depression detection. These findings suggest that such an automated process can improve the detection of patients with suspected depression.

4.3 Neurodegenerative Disorders

Neurodegenerative disorder (ND) is a brain disease caused by the damages affect the neurons' synapses. It's a progressive and irreversible disease, characterised by losing cognitive functionality. Subjects with dementia suffer from memory loss, weak communication skills, mood changes, depression and decline in the understanding [213]. At advanced stages, the patient cannot speak and perform simple daily life functions [214]. Many diseases can lead to ND, such as Dementia with Lewy bodies (DLB), Frontal Lobe Dementia (FLD) and Vascular Dementia (VD). However, Alzheimer disease (AD) is the most common cause of dementia, with an average of 60-80 % of dementia's [215]. AD results from two abnormal fragments of a protein called Tangles Plaques accumulated in the brain neuron's synapsis. This destroys the connections between the neuron cells that contain the sensation, motor skills and memories. Therefore, memory impairment is the most common AD symptom, affecting executive functions and the individual's behaviour [216]. VD is the second most common form of ND with 20% ofND cases. This kind of ND caused by blood circulation damage in the brain may result from small blood clots obstructing brain cells' oxygen supply. Symptoms of VD depends on the affected part of the brain, and may include: language, reading, communications and writing [217]. DLB is another type of ND with at least 10-15% cases having movement and control issues. This type of ND caused by abnormal protein clusters called (Lewy bodies) assembled in different parts of the brain. DBL symptoms include sleep disturbances, swallowing problems and hallucination. Additionally, there are several conditions were subjects show similar memory problems to that caused by ND [218]. These conditions have the chance to be cured compared to conditions caused by neurodegenerative diseases (ND). However, diagnosing the disease's actual cause is challenging due to the overlapping symptoms and the lack of suitable screening biomarkers. The mild cognitive impairment (MCI) is repeatedly mentioned as an early course of dementia; however, the rate of MCI patients developing AD conditions is less than 15% [219]. Most MCI conditions are caused by several factors, such as, drugs or depression, which can be controlled and thus conditions can be improved [218]. Functional Memory Disorder (FMD) is another type of clinical conditions where patients have memory problems without objective cognitive deficits caused by a stressful event or psychological issues. Considerable pressure on secondary care memory services resulted from a large increase of referrals from primary care for people with memory complaints [220]. The attempts to find early diagnostic tools has led to an over 600% increase in referrals to memory clinics in the UK during the last 10 years and resulted in a significant pressure on health care system [220]. These changes have increased the number of identified ND patients. Additionally, many patients referred to specialist memory clinics have Functional (non-progressive) Memory Disorder (FMD) concerns without objective evidence of cognitive deficits. Hence, improvements in screening procedures would be strongly desirable and could lessen the pressure on limited health care resources [221]. Typical memory assessment begins with a conversation with a specialist where patients are asked a series of memory-related questions. This interaction provides useful insights into the cognitive state of the patient. Therefore, reliable, non-invasive and automated diagnosis tool is urgently required [221].

4.3.1 Neurodegenerative Disorders diagnosis

ND diagnosis includes several steps that require examination and specific tests which can be lengthy and complicated. The first step is to investigate the presence of co-morbid disease, which show similar symptoms to dementia. Next, memory, orientation, attention, language and executive function will be evaluated in a series of cognitive tests. More neurological tests may be required, such as gate and cranial nerve examinations [222]. Additionally, several biological markers can be used to assess the diagnosis process such as Cerebro Spinal Fluid (CSF), Electroencephalogram (EEG), Computed Tomography (CT) scans, blood tests, Positron Emission Tomography (PET), Single Photon Emission Computed Tomography (SPECT) and Magnetic Resonance. Imaging (MRI). These examinations could be costly and cumbersome and expose patients to high radiation [223]. Moreover, several behavioural changes are considered by neurologists during the diagnosis process such as walking, communications and sleeping pattern taken into consideration. To assess these, several tests exist to help examine speech and language capabilities of the individual's [224].

4.3.2 Cognitive state diagnosis tools

Tools for cognitive state diagnosis were developed to assess the patient's performance in a range of cognitive domains including memory, time orientation, language, verbal fluency and visuospatial abilities. The examiner will compare the patient's final test score against the cut-off to evaluate its condition. Next sections will briefly introduce the most employed tools for dementia diagnosis. However, these tools are limited in term of sensitivity (true negative rate) or specificity (true positive rate).

Minimal Mental Status Examination

The Mini-Mental State Examination (MMSE) is widely used in clinical settings to examine the cognitive impairment with a period of (5-10) minutes administration time. This tool developed by Folstein et al. [225]. Moreover, it is used to diagnose dementia and estimate the severity of cognitive impairment. MMSE estimates the severity of cognitive impairment using five categories: registration, language, memory, attention and orientation. There are 11 questions in this test with the highest score of 30 points (regarded as a healthy cognitive state). A score below 10 indicates a major cognitive decline, while mild cognitive ranged between (20-25) and moderate impairment scores lie (20-10). However, MMSE is sensitive against age, years of education and events of the gradual changes that occur to AD patients [226]. Additionally, MMSE shows low success rate to diagnose mild AD from healthy patients.

Montreal Cognitive Assessment

The Montreal Cognitive Assessment (MoCA) is another cognitive impairment evaluating tool used extensively [227]. Ziad Nasreddine presented this tool in Montreal, and the administration time is 10 minutes. The top score is 30, and the cut off for normal state is 26. MoCA tests several cognitive areas, including language, memory, orientation and visuospatial. Compared to MMSE, MoCA has a higher success rate for diagnosing MCI compared to MMSE. However, the specificity is insufficient.

4.4 Neurodegenerative cognitive declines assessment

This experiment presents a novel automatic system based on visual features to identify individuals at high risk of developing dementia. The feasibility of using the visual features based on eye blinks, eye-gaze and head movements investigated to differentiate between patients presenting to specialist services with memory problems related to progressive ND, MCI or FMD. Dataset used in this experiment provided by Royal Hallamshire hospital, which includes video recordings of individuals while answering questions posed by an intelligent virtual agent (IVA). A huge increase in referrals cases to secondary care memory services from primary care for people with memory complaints. This resulted in considerable pressure on diagnostic pathways [228]. The effort to find early diagnostic procedure caused an over 600% increase in referrals to secondary care memory clinics in the UK during the last ten years. Further, it challenged the diagnostic pathways [228]. Although these dramatic changes have raised the identification of patients with ND, a large rate of the patients referred to specialist memory clinics have FMD without objective evidence of cognitive deficits. Moreover, Memory clinics in the UK are experiencing an increasing number of referrals [224] and are struggling to meet the demand for specialist diagnostic services [229]. Therefore, improvements to screening procedures would be highly desirable and could better manage limited health care resources [221]. The current diagnostic process focuses on clinicians interpretation of the history given by patients and their companions. Neuropsychological tests complement the clinical impression formed in this way either in the form of simple short screening tests or formal neuropsychological examination over several hours, and brain scanning (Magnetic Resonance Imaging (MRI) or computerised tomography(CT) [221]. This type of interaction helps the clinician to get important insights into the cognitive state of the patient. The clinician will note the patient response to the questions, whether answers are quick and extensive or brief and incomplete. Considering this process's cost and complexity, noninvasive and reliable diagnosis tools that are automated can be reused and available remotely urgently needed [221]. Recently, eye blink rate used by [230] as a biomarker for MCI diagnosis. Eye blinks were monitored using two gold skin electrodes above and below the left eye while subjects cognitive function was assessed using a battery of neuropsychological tests. This research shows that eye blink rate in participants with MCI was significantly higher than in healthy controls. However, this study has several limitations that need to be resolved, such as the use of horizontal and vertical gold skin electrodes placed above and below the left eye, which can be considered invasive. Mirheidari et al. [231–233] proposed an automated approach for the differentiation patients with cognitive complaints due to ND or FMD using features extracted from Conversation Analysis (CA) [224, 234]. A set of linguistic, acoustic and visual-conceptual features were extracted and used to train a group of classifiers in their work. The highest classification accuracy of 97% was achieved using a linear support vector model. The approach proposed here is different from the previous studies by Mirheidari et al. in terms of complexity and features used. Here, the method explores visual-only features extracted directly from patients' video recordings. The previous study depends on more complex features based on clinicians and carers' contributions to the interaction, natural language processing, and automatic speech recognition.

4.5 Neurodegenerative disorder Dataset

The dataset used in this study using collected as part of a study conducted in the neurology memory clinic at the Royal Hallamshire Hospital in Sheffield, United Kingdom. The study was approved by NRES Committee Yorkshire The Humber - South Yorkshire. This study aimed to capture audio-visual recordings of people answering similar style questions to those typically used in diagnostic neurologist-patient conversations. The memory-probing questions were put together by a multidisciplinary team involving neurologists, linguists and experts in conversation analysis. Participants received an information sheet before taking part in the study and allowed to ask questions. Patients agreed their data be used for additional analyses by the research team, but not their recordings made publicly available. Patients were encouraged to answer eight questions posed by the IVA verbally and told that they could answer as briefly or extensively as they liked. In total, 24 participants took part and out of these, a total of 18 recordings of patients interacting with the IVA were further analysed (6 ND, 6 FMD and 6 MCI).

For the rest of the recordings, 4 subjects were excluded with depressive pseudodementia and 2 in whom the diagnosis was not clear. Patients referred because of memory complaints by a general practitioner or hospital consultant. Participants encouraged to bring a companion such as a carer or family member (if available) along to their appointment—more details about the procedure and the participant's selection procedure provided in [233]. At the memory clinic, patients were assessed by a neurologist specialising in diagnosing and treating memory disorders. Additionally, patients assessed by the Addenbrooke's Cognitive Examination-Revised (ACE-R) cognitive assessment [235]. Also, patients were examined for depression using PHQ-9 [55], where patients with high scores were excluded from this study [233]. The Generalised Anxiety Disorder (GAD7) questionnaire [236] also done by all the patients. The decision of ND or FMD were formulated by brain Magnetic Resonance Imaging (MRI) findings, consultant neurologists specialised in the treatment of cognitive disorders, MMSE [225]; tests of abstract reasoning [237] tests of attention and the executive function [238]; category and letter fluency; naming by confrontation and language comprehension [239]; and tests of short and long-term memory (verbal and non-verbal) [240]. Table 4.5 gives an overview of participants' details and test scores.

Table 4.5 Participants' details and test scores. ACE-R: Addenbrooke's Cognitive Examination-Revised; MMSE: Mini-mental state examination; PHQ9: Patient Health Questionnaire-9; GAD-7: Generalised Anxiety Assessment 7. Unpaired T-test was used. ns = not significant.

	FMD(n=15)	ND(n=15)	Cut off	Max score	P-value
Age	57.8 ±2.0	63.7 ±2.3	N/A	N/A	p =0.06
Female	60%	53%			ns*
ACE-R	93.0 ±1.4	58 ±58	88	100	p<0.0001
MMSE	28.9 ±0.2	18.8±2.0	26.3	30	p<0.0001
PHQ9	5.6 ±1.0	5.3 ±2.0	5	27	ns
GAD7	4.7 ±1.2	4.8 ±1.5	5	21	ns
History taking part in minutes	range (10.1-32.3)	range(7.3-29.0)			

4.6 Proposed method

The proposed method aims for an early detection tool for patients showing progressive ND-related cognitive problems using visual features in their video recordings. The proposed system pipeline is similar to the method introduced in section 4.1.5, and it formed of pre-processing, feature extraction and machine learning-based classification.

4.6.1 Pre-processing

The clinical assessment sessions were recorded using a "ZOOM H2N" portable digital recorder. The device placed on the table within 1 meter of neurologist, patient and accompanying person. Video recordings were produced in "MP4" format with 30 frames per second. Recordings were segmented according to the asked questions in each session. The intention behind this is to investigate the effectiveness of features extracted from each question in classifying the patients.

4.6.2 Feature extraction

In this study, the aim is to explore the potential of using the visual feature to differentiate between FMD, MCI and ND. Several visual features were extracted including head pose, eye gaze movements and eye blink.

4.6.3 Eye movement features extraction

Eye movement features

For eye blink features, the technique presented in chapter 3 employed to extract eye blink features. Eye gaze features extracted using the following steps:

- Face Detection and Alignment: The face detection module gets image frames from the webcam source then, zface [160] library is used to detect faces in each frame. Next, the face alignment module aligns a non-rigid facial feature point model to each detected face.

- Feature Extraction This module extracts features for each detected face, including facial landmarks and eye pupil position [241]. Iris localisation method is based on [242] method that tries to find the possible eye center in a given image. The method is given in equation 4.6.1, where C is the proposed eye center, g_i is the gradient vector at position xi, d_i is the normalised displacement vector. In a given eye image, the optimal eye center location is searched for using dot products between the normalised displacement vectors d_i and gradient vectors g_i as shown in Figure 4.4. Given that the scalar product is maximal when both vectors are co-linear, the center c can be found as the biggest dot product value.

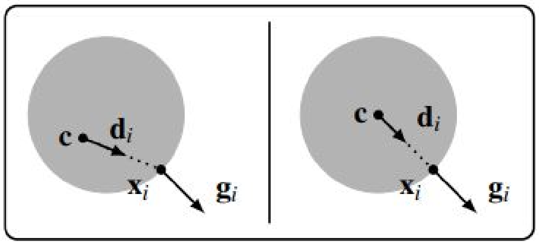


Figure 4.4 Unaligned and aligned pupil center and iris edge gradient vectors (left and right respectively adopted from[242].

$$C = argmax\frac{1}{N}\sum_{i=1}^{N}(d_i^T \cdot g_i)^2 \tag{4.6.1}$$

Algorithm 1 EYE CENTRE LOCALISATION

```
procedure $EYE\ (c, x)$
Let $c$ be the possible eye center
calculate $gi$ as the gradient vector at position $xi$
calculate the displacement vector $di$ from $C$ to $gi$

    for <each pixel in image> do
        <find N= $d_i$ ·$g_i$ >    end for
            return MAX(N)
        end procedure
```

Figure 4.5 Pseudo-code used for eye center localisation.

Head pose features

To assess head movement usability for the diagnosis of memory disorders, facial landmarks were selected that represent the face's rigid movement. Two groups of facial landmarks were investigated to quantify head movement. The first group contains landmarks for the nose region and includes landmarks *28,29,...,36*. The second group of landmarks form the face contour and includes landmarks *1,2,....,17*. The approach for extracting head movement features includes calculating the Euclidean distance between adjacent frames, which describes the change in *x,y,z* coordinates. Repeating this approach for all frames, a feature vector is generated describing the head movement velocity. Similarly, the second-order difference is applied to the obtain velocity feature that describes the head pose's acceleration profile. High confidence detected landmarks is used in this process to discard landmarks that detected with low confidence by the tracker. A summary of these facial landmarks is summarised in Table 4.6.

Table 4.6 Facial landmarks used for head movement features.

Feature	Facial landmarks
Face Contour	1,2,....,17
Nose	28,29,...,36
Nose + Face Contour	1,2,....,17,28,29,...,36

4.6.4 Validation Scheme

To ensure a robust performance evaluation and reduce overfitting, cross-validation technique is widely used in machine learning pipeline [243]. Cross-validation with 10 folds (k=10) utilised to obtain equal parts of data. The model trained using nine out of ten folds and tested with the last 10th fold. This step repeated until all data folds used in the training and testing process. However, this did not generate the validation set automatically. Instead, the nested cross-validation method utilised, which uses the outer k-fold cross-validation loop to split the data into training (9/10 data) and test folds (1/10 data). An inner loop is used to select the model via k-fold cross-validation on the training fold. Model hyper-parameter tuning and feature selection were investigated, and the model with the best features and parameters tested using the test folds. This procedure runs through all the loops, and the average of the best model scores across the outer test folds reported as the final result. The Scikit learn library was used to perform this task [244].

4.6.5 Results

The results achieved from this study suggest that the proposed machine learning model based on the analysis of visual features from subjects with memory complaints is capable of detecting differences between the three classes of ND, FMD and MCI. The discriminating potential of visual features explored using the Adaboost machine learning algorithm and tested using validation technique described in section 4.6.4. To determine each question's diagnostic contribution, visual features, i.e.(blink, head pose and eye gaze) were classified separately for each of the eight questions using the AdaBoost algorithm. Additionally, visual features were analysed by computing the averages across the whole given recordings. Then, z-score normalisation calculated, and the class-specific average histograms were produced. As shown in Figure 4.6, eye gaze for ND subjects increases in the right direction from 0 to 3. This might indicate head turnings towards the accompanying caregiver. This experiment's results are shown in Table 4.7, where two scenarios were tested: features from each question tested separately and features from a full video recording. For the first scenario, question 7 reached a maximum at 68.3%.

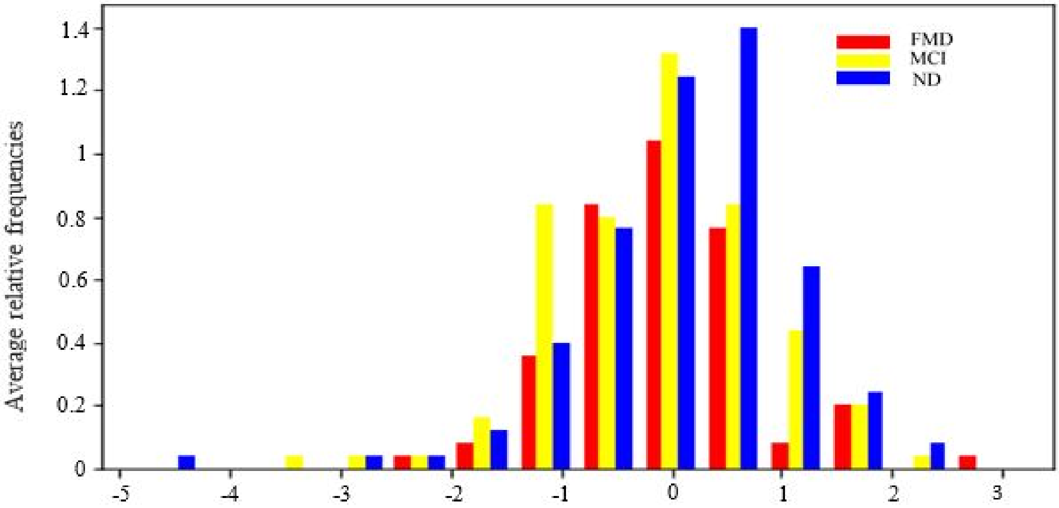


Figure 4.6 The histogram for gaze intensities for FMD,MCI and ND subjects.

Table 4.7 Classification results for MCI vs ND vs FMD subjects using Adaboost classifier.

Questions	Accuracy	F1 score	Precision	Recall
Q1	62.3%	0.62	0.63	0.62
Q2	52.55%	0.52	0.52	0.52
Q3	55.3%	0.55	0.59	0.55
Q4	59.5%	0.59	0.59	0.59
Q5	58.3%	0.58	0.58	0.58
Q6	55.55%	0.55	0.54	0.55
Q7	68.3%	0.68	0.67	0.68
Q8	61.2%	0.61	0.59	0.61
All questions	96%	0.70	0.69	0.70

4.6.6 Discussion

This study shows that using visual features could help as a complementary method in diagnosing patients presenting with memory concerns. This study aimed to develop a machine learning model that learns from visual recordings and can detect patients' cognitive status referred to a specialist memory clinic. In this study, a ternary classification was used; FMD, ND or MCI. The highest classification accuracy achieved was 96% using visual features from all questions asked during the diagnosis session. These features include head pose, eye gaze and eye blink features. A study done by [230] found an increased rate of eye blinks in MCI subjects. Similarly, [245] found that the eye movement abnormalities are an indicative feature for the assessment of memory disorders. Comparing this study to [233] which used the same dataset, the proposed model-based exclusively on visual features and requires fewer computation resources. In contrast, [233] is based on a combination of lexical, acoustic, visual-conceptual and semantic features. Moreover, the classification approach used by [233] included input features from a neurologist and accompanying persons, whereas the proposed model uses only visual features captured from video recordings. Using visual features from full video recordings, the overall classification accuracy improved to 96% in contrast to the complex model used by [232]. The sensitivity and specificity of the developed model were 92.30 and 99%, respectively. These results can be compared with other memory disorder screening modalities. For example, although Positron Emission Tomography (PET) is associated with higher sensitivity and specificity (86% for both), is invasive and need the injection of a radioactive tracer via a peripheral cannula and requiring that patients to be fasting for four hours before the test [246]. Single-photon Emission Computed Tomography (SPECT) is another screening tool capable of showing early changes of neurodegenerative disorders with high sensitivity (86.0%) and specificity (96.0%), but is costly and cumbersome as PET and exposes patients to a high radiation [223]. Comparing the proposed method to Mini-Mental State Examination (MMSE), although the MMSE shows high sensitivity (87.3%) and specificity (89.2%) scores, it is not effectively sensitive at the early stages of memory disorders, and it is affected by patient's education level [246]. The clock drawing test (CDT) screening shows high screening performance (sensitivity 92.8% and specificity 93.5%, however, it is not used frequently for memory testing which limits it is usefulness [247]. Moreover, the scoring having 8 different assessment setting which may be tricky to assess. Importantly, the described tools above are the state-of-the-art for neurodegenerative disorders screening and utilised widely. While the proposed method tested on a small dataset, performance needs to be validated with a large dataset. This study highlights the importance of developing a low-cost neurodegenerative disorder diagnostic method based on visual features. This could help in patients' assessment procedure with cognitive problems as they can be easily deployed due to low-cost hardware and less computational complexity. One limitation in this study that the dataset used is relatively small as it contains 6 cases in each group.

4.7 Summary

Results from this chapter lead to several conclusions. Firstly, visual features extracted from video recordings are important in diagnosing mental disorders, i.e. depression and neurodegenerative disorders. For depression diagnosis, video recordings from the AVEC2014 dataset used to build a machine learning model based on eye blink features. The results from the proposed model outperformed the baseline, and most of the studies used the same dataset in the literature. The model's performance was informative in predicting BDI and PHQ scores with better MAE and RMSE than the baseline scores in the AVEC2014 dataset. Classification accuracy was higher for the text reading task than for all test scenarios for the task's answering question. Results show that eye movement abnormality is in line with the psychology literature in that depressed subjects shows a general slowing of cognitive processes. These findings suggest that eye blink features can be deployed as an objective evaluation tool for depression, and these features can be obtained using low-cost platforms. Secondly, experiment 2 show that visual features extracted from head and eye movements significantly impact the differentiation between ND, MCI and FMD. These features are likely associated with changes in the neurobiology associated with a given neurodegenerative cognitive disorder and reflected in the clinicians' visual interaction. Moreover, the proposed approach can be deployed at clinics as it requires minimal effort. Finally, despite the limitations in this study, the findings demonstrate the visual-only features offer a potential alternative to complex and costly diagnosis tools. Therefore, it has the potential to be employed in the early stages to assess people with cognitive complaints.

Chapter 5

Bipolar disorder data exploration

5.1 Introduction

Bipolar disorder (BD) is a chronic condition characterised by mental state variation, episodes of depression and mania. In a manic state, BD patients become easily irritated, hyperactive, speaking loudly, and experience reduced sleep. Traditional procedures for bipolar diagnosis rely on self-assessment or clinically administrated questionnaires, which are subjective. Objective advance diagnosis and following medical care can positively influence the life quality of individuals who have bipolar disorder and help early access to the treatment. Therefore, an automated bipolar disorder diagnosis tool will benefit the patients and the health care system, as this disorder is permanent and needs continuous monitoring and treatment. For this aim, a novel modelling technique based on the extraction of indicative temporal feature trajectories (i.e. trajectories of emotional profiles, AUs, eye gaze, emotions and head pose) at the different time scales has been developed. This can be considered a valuable tool for mental health care as it can improve the diagnosis and tracking of bipolar disorder or relevant mental disorders. The proposed methodology is validated on the audio/visual dataset from the AVEC2018 bipolar corpus, which targets the prediction of mania levels in patients with bipolar disorder and is the only accessible dataset that contains audio and video recording for structured interviews of patients suffering from bipolar disease. This chapter presents the predictive analysis of visual features based on statistical analysis and machine learning.

5.2 Literature review

Emotion recognition recently turned out into an important research topic, as emotional information is essential for human communication. The Audio-Visual Emotion Challenge (AVEC), is a yearly challenge since 2011 which aims to further the progress of automatic continuous emotion recognition in the wild by applying new multimedia processing techniques and machine learning methods [120, 168, 196, 197, 248–251]. A framework of natural, spontaneous emotion recognition and a benchmark to evaluate the recognition techniques are supplied in each challenge. Since 2013, the challenges aim to explore depression prediction and recently, bipolar disorder prediction is introduced as a challenge in the AVEC2018. BD is a persistent (possibly lifelong) mental health disorder, with subjects experiencing occasional episodes of depression and mania lasting from days to months. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [15] describes bipolar disorder as a complicated mental disorder that changes an individuals mood to a large extent. Additionally, bipolar disorder affected subjects suffers from intervals of depression with episodes of mania characterised by elated mood, increased activity, over-confidence and impaired concentration. For example, a patient in a manic state shows increased energy, decreased demand for rest or sleep and interest in social activities. [252] [253]. BD has been ranked by the World Health Organisation (WHO) among the top 10 for young adults according to the diseases of disability-adjusted life year (DALY) indicator [254]. Remission rate and treatment acceptance are still low, despite improvements in bipolar disorder treatment. Treatment opposition is one of the big challenges for bipolar disorder [255]. This is mainly caused by the delay in diagnosis, frequent hospitalisations, and inflammation due to inadequate detection of depressive episodes. Turkish Audio-Visual Bipolar Disorder corpus was introduced by [256] as a subset of the dataset introduced in the AVEC2018. This dataset is annotated for the existence and severity of bipolar disorder using scores from the Young Mania Rating Scale (YMRS) [16] [120].

YMRS questionnaire has been developed by [16] to examine manic symptoms using 11 items. Four items in YMRS are weighted on a range between 0-8 assessing speech, aggressive behaviour and irritability. The rest of the items are weighted on a range between 0-4. This test can be administered by a psychiatrist or be self-reported. In the AVEC2018 dataset, YMRS test done under psychiatrists supervision [120]. The intensity of bipolar disorder is reported according to remission, hypo-mania, and mania states [256]. Where remission (YMRS 7) is the lowest intensity of the bipolar disorder and can be treated through medication [256] [120]. When the YMRS score ranging between 8 and 20, the hypo-mania state is recognised. Whereas YMRS score, greater than 20, is regarded to be in a state of mania which is considered as much higher intensity than hypo-mania and being abnormally energised mentally and physically. Early diagnosis and efficient treatment of this disorder can positively affect patient quality of life. Automated screening methods can help to achieve this goal. The automated BD diagnosis is understudied, and the AVEC 2018 is the first dataset to provide the records of structured interviews of individuals with different severity of the bipolar disorder. Multiple research groups have participated in the AVEC2018 and demonstrated a variety of modelling and inference approaches. Visual features information from facial landmarks and texture-based features extracted by [256], these features utilised using fine-tuned Convolutional Neural Network (CNN). Using audio modality,[256] employed Interspeech 2010 paralinguistic challenge features [257]. Finally, extreme learning machines [258] employed and partial least squares regression [259]. The authors in [260] used video sequence to described the displacement of the upper body points using Histogram of Displacement Range (HDR). They applied Deep Neural Network (DNN) and Random Forest for bipolar depression classification. Meanwhile, [261] investigated eye-gaze angle, head-pose, and only six facial action units using three classifiers, namely, SVM logistic regression and Greedy Ensembles of Weighted Extreme Learning Machines.

5.3 Data description

The dataset used in this chapter is supplied as a part of the AVEC2018 BD sub-challenge, which is a subset of the Turkish Audio-Visual BD Corpus [256]. This dataset was published at Affective Computing and Intelligent Interaction Conference (ACII Asia 2018). Participants had to classify audio-visual recordings of individuals into three classes, i.e. states of remission, hypo-mania, and mania according to the severity of the bipolar disorder as measured with the help of YMRS [16] according to the following rules :

1. Remission: YMRS ≤ 7

2. Hypo-mania: 7 < YMRS 20

3. Mania: YMRS > 20

A total of 50 patients (34 male and 16 female) aged 18-54 and 39 healthy controls (23 male and 16 female) aged 18 to 57 years are included in the dataset. The dataset contains 218 recordings in total, 104 recordings in the training partition, 60 in development partition and 54 for the test partition. Subjects are evaluated by performing seven tasks. These tasks range from explaining the activity's reason, describing happy and sad memories; counting up to thirty and explaining two emotion triggering pictures.The demographic information for the dataset used in this experiment shown in Table 5.1.

Table 5.1 Demographic information of the AVEC2018 BD dataset.

	Male	Female
Participants	23	11
Age range	(18-52)	(23-48)
Age mean	33.37	36.27
Total recordings	107	57

5.4 Visual features characteristic of bipolar disorder

The AVEC2018 aims to predict the intensity of mania for individuals with bipolar disease. As the labels are provided depending on YMRS Score [16], the current research investigated the core symptoms described in the manic phase of bipolar affective disorder for individuals with mania following the YMRS. Therefore, the research focused on identifying key features characterising bipolar disorder, contradicting brute-force approaches that ignore background knowledge about the condition. Inspired by the YMRS questionnaire, The investigation of the elevated mood, irritability, and motor activity variations during the speech were proposed. During the bipolar phase, subjects exhibit specific emotional profiles, characterised by increased anger and irritation. To extract such emotional features, a facial emotion recognition model developed to obtain emotional profiles for each image in the dataset videos using transfer learning technique (section 5.7). In addition to these transfer learning features, the presence and intensity of 18 key AUs were extracted from each video frame in the given video recordings. Irritability or dull state can be obtained from eye features which provides information about the motor activity. Eight statistic features that describe subject eye gaze directions were extracted for left and right eyes separately. Three-dimensional head pose (i.e. yaw, pitch and roll) is an important feature that describes increased energy and movements. These features were computed using OpenFace toolkit [84], which provides a comprehensive set of tools for image pre-processing and visual feature extraction. These features include eye-gaze, facial AUs, facial landmarks, a histogram of oriented gradient (HOG), and head pose [123]. For audio modality, OpenSmile toolkit [262] was used to extract Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features [263] which have shown its usefulness in representing information related to paralinguistics, prosody and emotions from speech acoustics. These features are closely related to the mental health of observed subjects in the AVEC2018 dataset. The aim is to employ the extracted features to investigate their extent's temporal trajectories at different time scales. This chapter investigated these using the standard machine learning pipeline, including data pre-processing, feature extraction, feature selection, model selection, and evaluation.

5.5 Characteristic feature distributions

To characterise the differences between the mania classes, visualisations of class-specific feature distributions were obtained. These were produced by first computing feature averages across the whole given recordings, then applying z-score normalisation to them, and finally calculating the class-specific average histograms for each considered mania level. Figure 5.1, 5.2 and Figure 5.3 shows examples of such distributions for a number of selected features. Figure 5.1, the distribution of head pose (yaw) is shown for remission, hypo-mania and mania groups. It can be noticed that subjects in mania group are more likely to move their heads towards the right direction than subjects in the remission group. This might indicate that subjects in mania group turn away from the screen more often than subjects in the remission group. This finding is supported by [264] who provided neurophysiological evidence for abnormal gaze processing in bipolar disorder and suggested dysfunctional processing of direct eye contact as a prominent characteristic of this condition. Figure 5.2, shows that subjects in mania group have a significantly higher intensity of AU15 (lip corner depressor) activation compared to people in the remission group. For eye-gaze features, Figure 5.3 shows that mania subjects more likely to move their eyes from left to right. The obtained distributions show that these features might provide high discriminative power for subjects' classification in the AVEC2018 dataset. In the next section, these features will be further assessed using several machine learning algorithms.

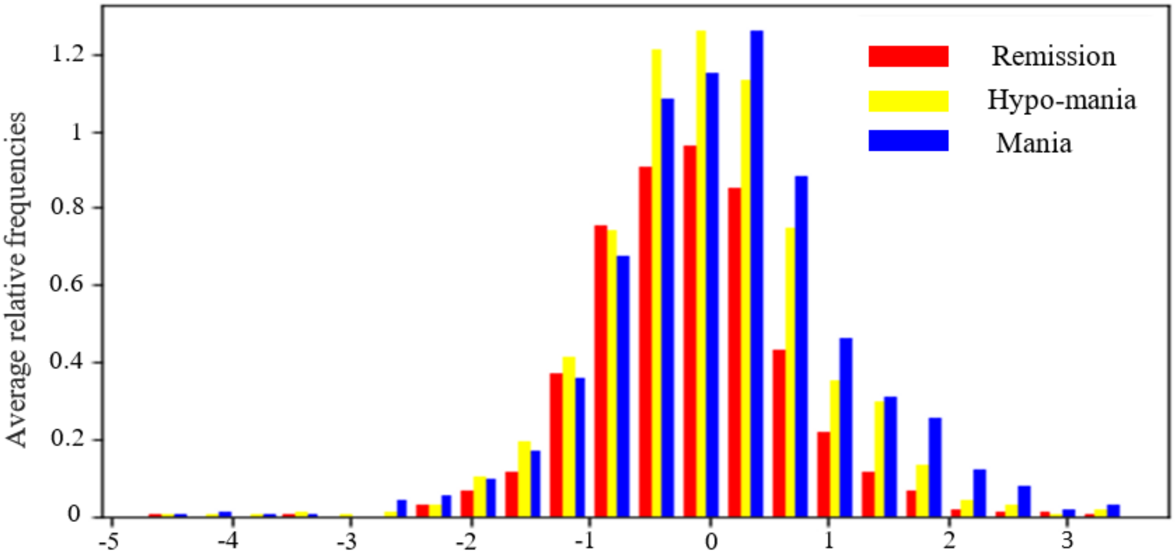


Figure 5.1 The histograms of head pose (yaw) feature for the three classes of AVEC 2018 dataset.

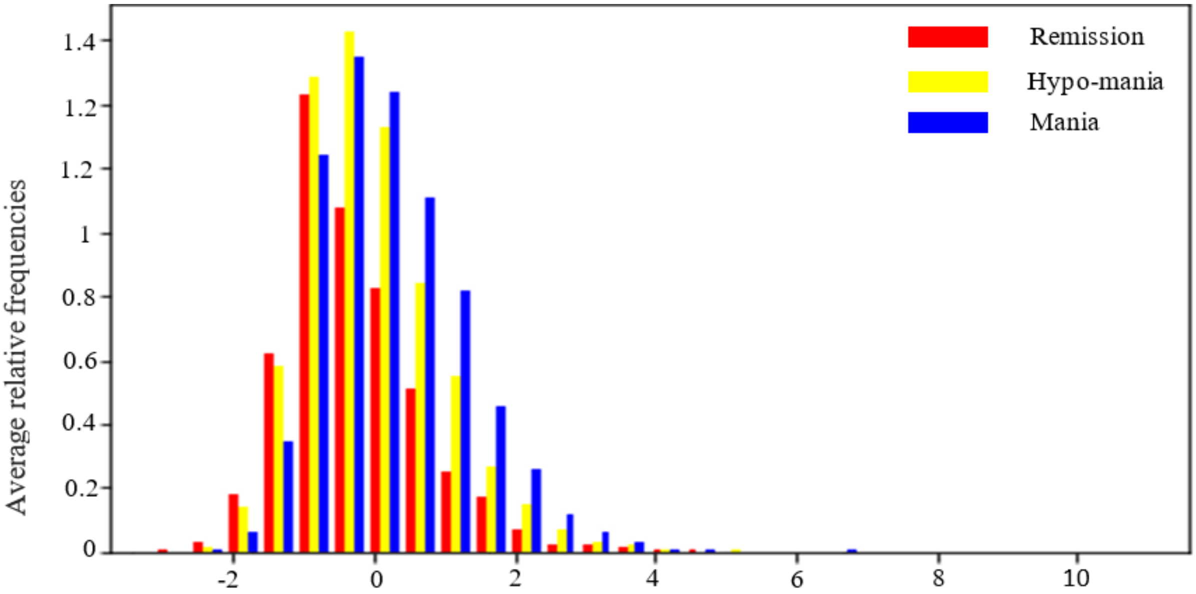


Figure 5.2 The histograms for lip corner depressor AU15 intensities for the three classes of AVEC 2018 dataset.

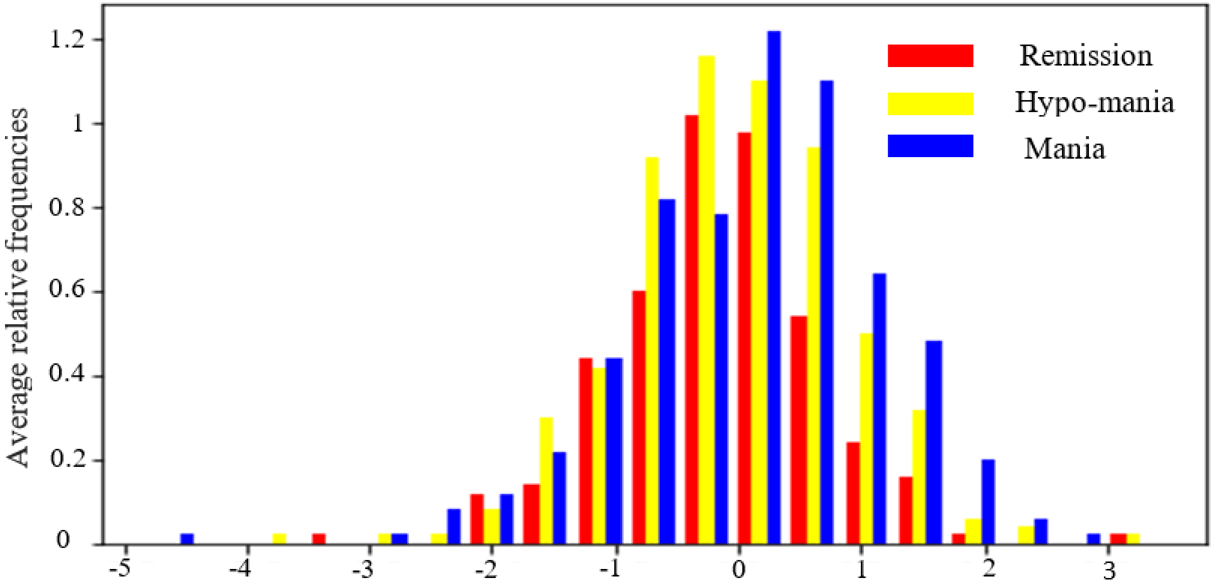


Figure 5.3 The histograms for eye gaze intensities for the three classes of AVEC 2018 dataset.

5.6 Feature selection and classifier evaluation

Considering the limited amount of observations in the AVEC2018 and the multitude of the considered visual features, selecting a smaller subset of strong predictors for model training and evaluation is necessary to achieve optimal performance. The feature selection was implemented based on Extremely randomised trees feature importance [265]. A subset of features is ranked based on their importance for the given task; then the best subset is selected. As shown in Figure 4.2 in chapter 4, each iteration consists of three nested loops; the outer loop performs a hyper-parameter search for feature selection. The two inner iterations are concerned with searching for the optimal parameters for the estimator. In each iteration, hyper-parameters are adjusted for feature selection which results in a new feature subset. Using this subset, another hyper-parameter search is done to search for the best parameter for the estimator. This pipeline design intends to optimise the model parameters to achieve the best results for those selected feature subset. Features and model evaluation results from this experiment are utilised in chapter 6.

5.6.1 Classification and regression analysis

This section describes several experiments which aim to investigate the effectiveness of the extracted visual features, i.e. eye gaze, head pose and AUs for the AVEC2018 bipolar disorder classification and regression tasks. For the classification task, the objective is to classify individuals into states of remission, hypo-mania and mania. In the regression scenario, the task is to predict patients total YMRS score, represented by an integer value in the range between 0 and 40. In each experiment, visual feature types (AUs, head pose, eye gaze) are employed individually and combined produce unweighted average recall (UAR) metrics for the considered classifier types. The same procedure was applied to the regression task. Classifier evaluation was conducted by splitting the dataset into train and validation partitions, as suggested in the AVEC2018. In this analysis, two feature extraction and temporal summaries were considered from full recording lengths and overlapping 9-minute long time windows. The latter was done to support the work on a deep learning model described in chapter 6, where observation lengths have to be standardised.

5.6.2 Model evaluation on full length recordings

This experiment is based on using visual features extracted from full video length. The sequence of this analysis follows the pipeline presented in section 5.6. The performance of each separate feature group and combined features are evaluated, to assess the contribution of each feature individually and when included in the selected feature subset. Finding out which feature shows a better performance might highlight important visual cues of bipolar disorder. Some interesting findings resulted from this experiment are presented in Table 5.2. Eye gaze features resulted in a higher Unweighted Average Recall (UAR) than other considered visual features. This suggests that eye gaze dynamics during speech are especially indicative of the extent of bipolar disorder. Additionally, features presented in Table 5.3 are concatenated together to enhance prediction on the development partition. This shows a similar trend in the accuracy as shown in Table 5.2. Similarly, Table 5.4 shows a regression analysis for each feature type separately and concatenated together.

Table 5.2 Unweighted Average Recall (UAR%) for ternary classification on development set using visual features form full recording.

Features	Random forest	Bagging	Adaboost
Eye gaze	41.66%	65%	63%
Head pose	35%	48.33%	59.25%
AUs	35%	60.0%	58.66%
AUs,Pose,Eye gaze	48.5 %	66.33%	62.5%

Table 5.3 Unweighted Average Recall (UAR%) using features that produced highest classification results from Table 5.2

Features	Random forest	Bagging	Adaboost
AUs,Pose,Eye gaze	0.5 %	62.33%	65.5%

Table 5.4 Mean Absolute Error (MAE) on development set using visual features form full recording as per YMRS scores.

Features	Random forest	Bagging	Adaboost
Eye gaze	11.3	8.9	8.2
Head pose	11.2	8.7	8.5
AUs	11.5	9.6	9.2
AUs,Pose,Eye gaze	11.4	10.2	8.1

5.6.3 Model evaluation on 9 minute time windows

In this experiment, data samples are prepared using 9 minute time windows for each video recording. Recordings less than this period are padded with zeros, while longer recordings are sliced into equal 9 minute long segments. Statistical summaries were then extracted from each time series and normalised using a robust scaling method (see section 4.1.4) using scikit-learn learn machine learning library [266]. Results from this experiment are shown in Table 5.5 below. These results need to be projected on full-length recording features to evaluate the classification accuracy. This prediction produced results comparable to the full recording experiment as shown in Table 5.6. Regression results for 9 minutes time-window and the projection on full-length video recordings are shown in Table 5.7 and Table 5.8.

Table 5.5 Unweighted Average Recall (UAR%) for ternary classification on development set using 9 minute time windows.

Features	Random forest	Bagging	Adaboost
Eye gaze	41.18%	53.08%	65.3%
Head pose	41.18%	55.33%	63.5%
Facial AU	47.36%	56.29%	61.32%
AUs,Pose,Eye	50 %	60.33%	62.5%

Table 5.6 Predicting unweighted Average Recall (UAR%) on full video length using 9 minute time windows model.

Features	Random forest	Bagging	Adaboost
Eye gaze	39.5%	50.23%	63%
Head pose	42.25%	54.53%	61.33%
AUs	48.86%	54.40%	59.48%
AUs,Pose,Eye	48.36 %	58.33%	61%

Table 5.7 Mean Absolute Error (MAE) on development set using 9 minute time windows as per YMRS scores.

Features	Random forest	Bagging	Adaboost
Eye gaze	11.3	8.9	8.2
Head pose	11.2	8.7	8.5
AUs	11.5	9.6	9.2
AUs,Pose,Eye gaze	11.4	10.2	8.1

Table 5.8 Predicting mean Absolute Error (MAE) for full video length using 9 minute time windows model.

Features	Random forest	Bagging	Adaboost
Eye gaze	12.8	11.23	9.4
Head pose	11.9	9.8	9.7
AUs	12.3	10.6	9.6
AUs,Pose,Eye gaze	10.2	9.8	8.9

5.7 Transfer learning for automated BD diagnosis

This experiment aims to develop and train a deep learning model for Facial Expression Recognition (FER) with fixed categories (e,g, neutral, sad, contempt, happy, surprise, fear and angry) using static images. This model is then utilised as a transfer learning tool to obtain emotional profile estimates for each image in the AVEC2018 video recordings. Transfer learning utilises knowledge gained from a deep learning task to solve a related problem. The transfer-ability of features increases as the used dataset are related for both tasks [267]. Inspired by the Visual Geometry Group (VGG-Face) model [268], three different network architectures (shallow, semi-shallow and deep) have been designed and adjusted to fit the image spatial dimensions and compared by their performance using FER2013 dataset[45].

5.7.1 Data pre-processing

Preparing quality data for different data science tasks such as pattern recognition and machine learning is a crucial requirement which impacts the system performance [269].

Applying data pre-processing techniques can help to speed up training and enhance the performance of CNNs. For this task, data samples must have the same spatial dimensions and normalised between 0 and 1 by dividing each image pixel by 255 in image samples. Data standardisation is another technique that can be done by calculating all images' mean and subtracting this value with all the images. The new mean of all images is now equal to zero. The images are then divided by the standard deviation of all images. Data augmentation is another technique that can be applied to get better performance from CNNs. This can increase the quantity of data, provide a type of regularisation and have more data variation. This technique can be can reduce the overfitting and enhance the network performance. Common data augmentation techniques are :

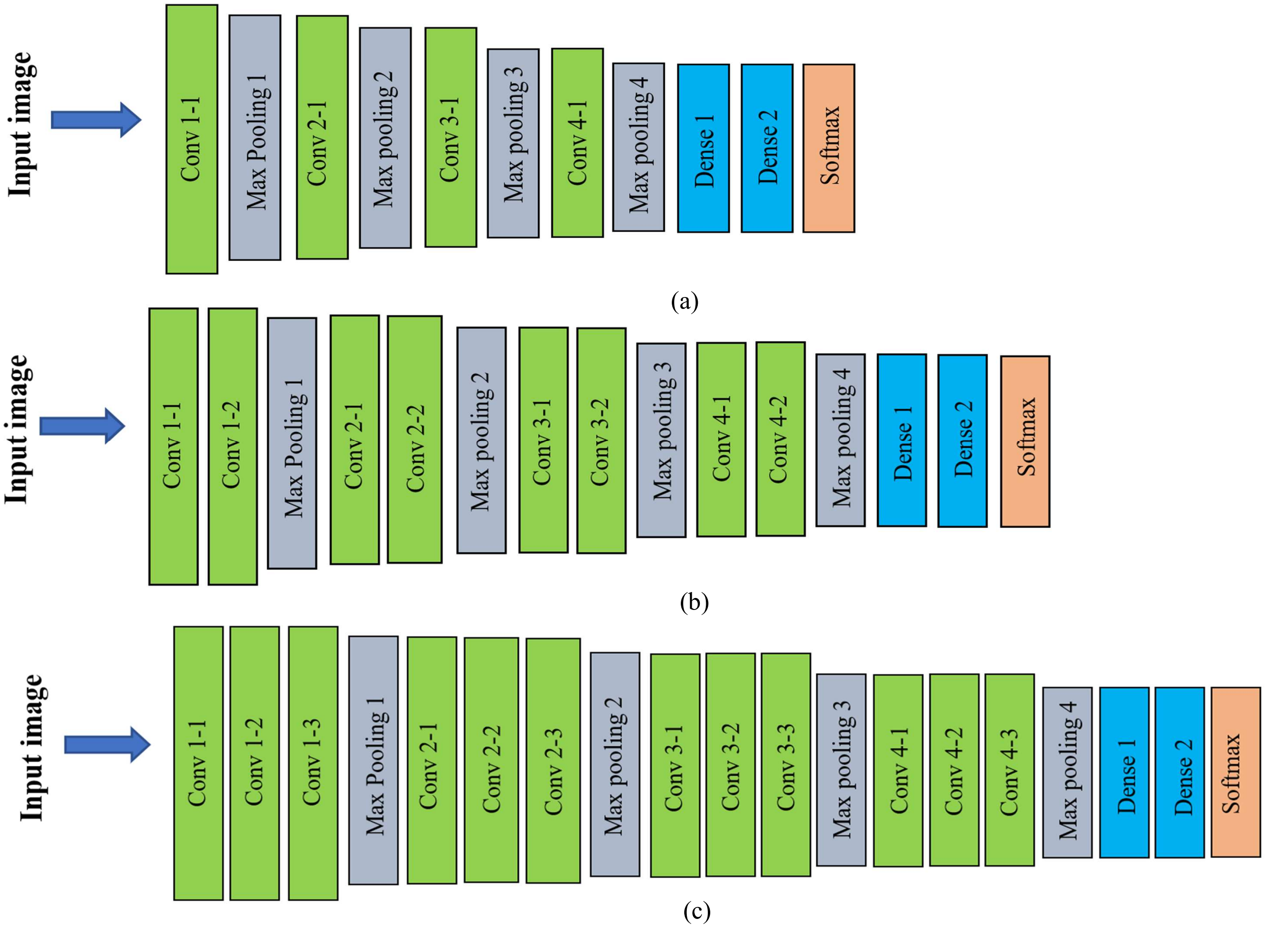
- Flipping : Flip an image horizontally or vertically to provide a different sample.
- Rotation : Rotate an image from the center point with an angle from 0 to 360.
- Stretching : Stretch an image vertically of horizontally.
- Rescaling : Rescale an image with the spatial dimensions, this is to provide a different size of the image within it is given initial dimensions.

In this experiment, training and test images for the experimented models were resized to 48x48 pixels and converted to grayscale. Then, images are normalised between 0 and 1, and the list above techniques are used for data augmentation technique.

5.7.2 Deep learning models for facial emotion recognition

Data samples are pre-processed according to the techniques mentioned in the previous section. As shown in Figure 5.4, the feature extraction block in each architecture is represented by multiple stacks of ReLU-activated 2D convolutional layers with 3x3 kernels alternated by max-pooling operators for downsampling. Such structure produces a hierarchy of features at the various spatial scales. The fully connected layers perform feature selection and classification in every given model variation on top of the model. This analysis's model variations differ mainly by the number of stacked convolutional layers between the downsampling operators.

The deep architecture features 3 Conv2D layers between max pooling and higher computational complexity, while the shallow network has only one convolutional layer between downsampling steps and can be trained much faster.



(c)

Figure 5.4 Network architectures for FER recognition. (a) shallow network, (b) semi-shallow network and (c) deep network.

5.7.3 FER model selection

To compare the three designed architectures' performance, the depth of the network and hyper-parameter tuning for each network is investigated to achieve the best model performance to compare the three designed architectures' performance. Data is split into training and test partitions, where the network is trained on the training set and validated by the test set. The model training and validation tests investigate the capabilities of shallow, semi-shallow and deep neural networks for FER task. Each model evaluation is implemented using 10-fold cross-validation. Every cross-validation fold includes 90% training data and 10% testing data. In such a model evaluation scheme, the semi-shallow model has achieved better accuracy than the other variants. Figure 5.5 shows the semi-shallow network's performance on the validation and test data. It is clear from the figure that the validation accuracy increases gradually after every epoch, while the test becomes nearly stable after 10 epochs. Figure 5.6 shows the model loss for training and validation. Both plots show an indication of overfitting as the improvement of the training and testing stops. This is because of the limited data provided. Therefore, the network cannot improve learning accuracy or provide useful information for new test samples to get the best model weights. Early stopping method is used once the model performance stops improving on a hold out validation dataset. The performance of the three different network architectures is shown in Table 5.9. Based on these results, the model produced using the semi-shallow network is selected as a transfer learning model. The best network weights are saved for transfer learning feature extraction, employed in the next sections. Transfer learning features are produced by applying the FER-trained model on images from the AVEC2018 videos and extracting the last dense layer's outputs before the softmax layer obtaining the 1x256 feature vector for each image the AVEC2018 dataset.

Table 5.9 FER recognition accuracy for the developed deep Networks using 10-fold cross-validation

Network size	Training Accuracy	Test Accuracy
Shallow	50.22%	57.80%
Semi-shallow	80.05%	63.80%
Deep	93%	61.22%

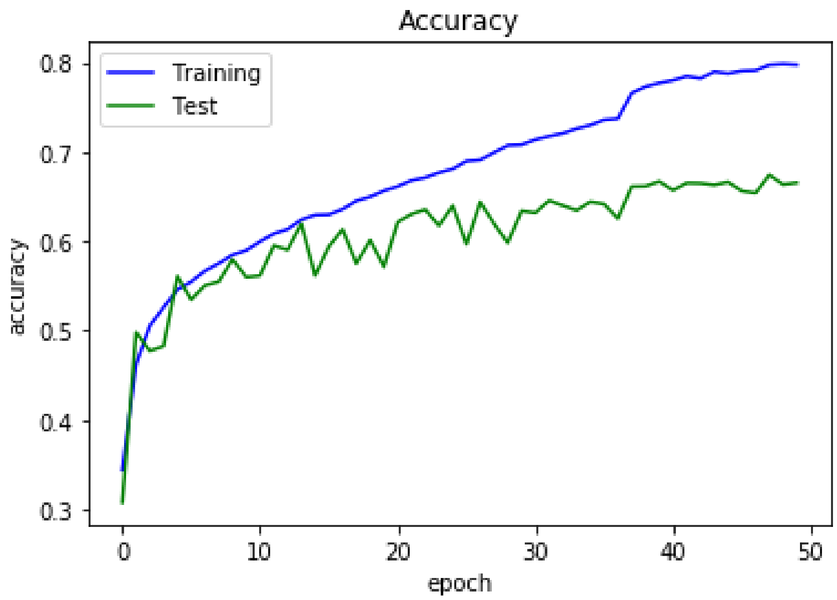


Figure 5.5 Loss function of the training process for the shallow network.

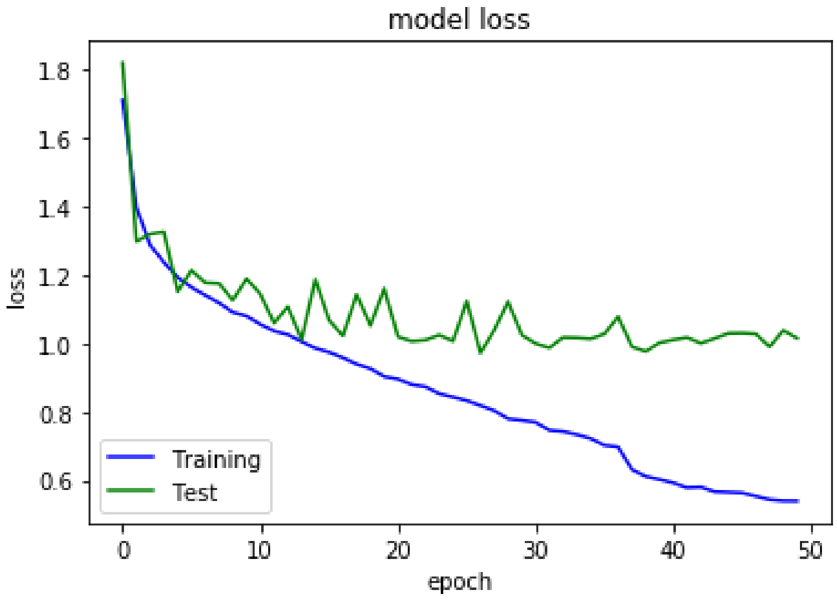


Figure 5.6 The loss function of the training process for the shallow network.

5.7.4 The application of transfer learning features to BD diagnosis

To utilise from the transfer learning technique, two pre-trained models were investigated. The first model is the semi-shallow model developed in the last section and the second model is the Visual Geometry Group (VGG) model developed for face recognition. The AVEC2018 dataset images were prepared by extracting images from each video frame, adjusting their sizes to fit the investigated models' requirements. Details of the tested models are given below :

- FER model transfer learning: facial images are extracted and aligned for each video frame in the AVEC2018 dataset, then the weights from the experiment in section 5.7.3 are used to extract pre-trained facial emotion features from the layer before the fully connected layer which produces feature matrix of size 1×256. These features were combined into an effective ($n \times d$) time series representation to describe the video at each frame-level where n is the number of video frames and d is the learned features. Considering that features at each column Di represent a specific emotion, the dynamic change of emotions are computed over time to extract several statistical features. The functions used to extract statistical descriptors are minimum, maximum, median, average, standard deviation, skewness and kurtosis. Python implementation was employed for extracting all the statistical descriptors.

- VGG-face: Visual Geometry Group (VGG), is a university of oxford, UK, a research group that have designed several well-known neural network architectures for object and face detection such as VGG16, VGG19 and VGG-Face [101] [268]. These networks have different layer size and speed in processing the data (e.g. 37 layers for VGG-Face, 16 layers for VGG16). In theory, VGG-Face is more suitable for FER as it was trained for the recognition of 2622 faces and shown state-of-the-art performance on several databases for face recognition[270]. The required input image size is (224×224×3) processed through a stack of convolutional layers, where the kernel size for each of the convolution layers is 3×3 with 1 padding and strides of 1. The fully connected layer handles the connection of features produced from the early convolution layers. These include edges, blobs and facial parts. Image features such as edges, blobs and facial parts are learned in the early layers separately. Fully connected layers handle the interconnectivity of these features to produce a response. The feature vector of 4096 dimensions used to produce a deep representation of the given image sequences extracted from video frames.

As shown in Figure 5.7 features extracted from both models are processed using the machine learning pipeline produced in Figure 4.2 in chapter 4. Features were investigated using full recordings and 9-minute time windows. Table 5.10 shows the results obtained using features from the full-length video recording. These features were then combined with AUs, pose and eye features to test whether it can produce any classification improvement.

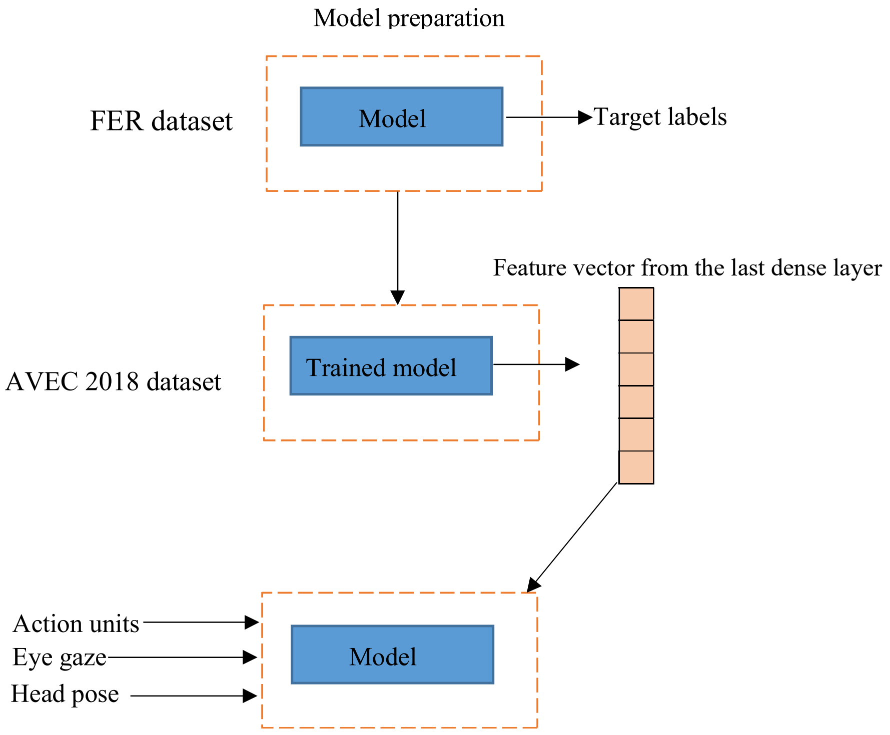


Figure 5.7 Applying transfer learning technique on the AVEC2018 dataset using FER model.

Table 5.10 Classification of FER model, VGG-Face, AUs, pose and eye features on full recordings.

Feature	Random forest	Bagging	Adaboost
Facial emotions	36%	61.33%	61.66%
VGG features	35%	50%	62.66%
VGG features,AUs,Pose,Eye,	35%	50%	63.66%

5.8 Summary

This chapter investigated the AVEC2018 dataset using machine learning algorithms and statistical analysis for the diagnosis of BD. Visual features were extracted according to the description of BD in the YMRS test. Additionally, a deep learning model for facial emotion recognition was developed. This model is used to detect emotion from video recordings by applying transfer learning technique. Furthermore, image features from BD patients recordings were extracted using VGG-face model. Classification and regression experiments were conducted using visual features from full recordings and 9-minute time window lengths. The purpose of these experiments was to assess the extracted features' capability to classify individuals into remission, hypo-mania and mania classes and predict patients YMRS score. The aim is to get the most descriptive features from the huge features vector and save machine learning model predictions to be used in the next chapter experiments.

Chapter 6

Deep learning for automated micro-expression modelling

In the previous chapter, the visual features extracted from video recordings of BD patients has been investigated using machine learning algorithms and statistical analysis. Useful features were obtained together with predictions for classification and regression tasks. These features' temporal trajectories were encoded using a set of statistical functionals applied to the whole recording duration, omitting short-scale sequences that represent facial micro-expressions indicative of the subject mental state. This limitation can be addressed using the deep learning paradigm, which offers automated feature design capabilities and representation learning at multiple time scales. This chapter is dedicated to the design of convolutional neural network architecture for automatic extraction of behavioural audio/visual features during the speech. By assuming that patients who share common behavioural patterns are likely to show similar physiological features, the comprehensive modelling of behaviours observable in data may produce indicative features for bipolar disorder diagnosis. This type of complex analysis is possible with the recent advances in artificial neural networks, which has provided research communities with new tools for data-driven automatic feature extraction from the available observations. This chapter describes multiple novel deep learning architectures that can capture episodic visual events commonly occurring in mental disorders. Several experiments were conducted exploring different deep learning architectures inspired by ResNet [271] and GoogLeNet [272]. In addition to that, a novel network output structure for simultaneous multi-purpose inference is introduced in the current chapter. The proposed network architecture has achieved evaluation accuracy results comparable to the state-of-the-art submissions for the AVEC2014 and the AVEC2018 challenges. The rest of the chapter is organised as follows: Section 6.1 describes data preparation for the AVEC2018 data. Section 6.2 provides an overview of the proposed deep learning architecture is introduced. Section 6.3 explains the output of the proposed deep learning architecture. Section 6.4 investigates several deep learning architectures for temporal feature extraction. Section 6.5 tests the proposed deep learning architecture on the AVEC2014 dataset. Section 6.6 presents the results and compares them with the literature. Finally, section 6.7 shows the discussion.

6.1 Data preparation

Time series analysis with Artificial Nural Networks (ANNs) asserts two major requirements for data pre-processing. Firstly, the input time series must be normalised to [-1, 1] or [0, 1] range, where non-linear activation functions commonly operate. Considering the amount of outlier data points in time series, a robust scaling method was selected. This method obtains normalisation statistics from an interquartile range of time series. Therefore, outliers in data do not contribute to the mean and standard deviation values used for standard normalisation. Secondly, time-series epoch lengths must be standardised to allow for inter-subject data analysis. This was achieved by segmenting the original videos into overlapping time windows of fixed length and zero-padding recordings shorter than the defined time window. To prohibit the zero-padded areas from affecting the feature extraction, a binary mask was obtained for each observation, with values of 0 marking the zero-padded region and values of 1 otherwise. Thus, input time series of different feature types (AUs, gaze, pose, etc) are represented by tensors $D \in \mathbb{R}^{T \times N_f}$ and a mask $M \in \mathbb{R}^{T \times 1}$, where T is the time window length and N_f is time-series depth (number of features per timestep).

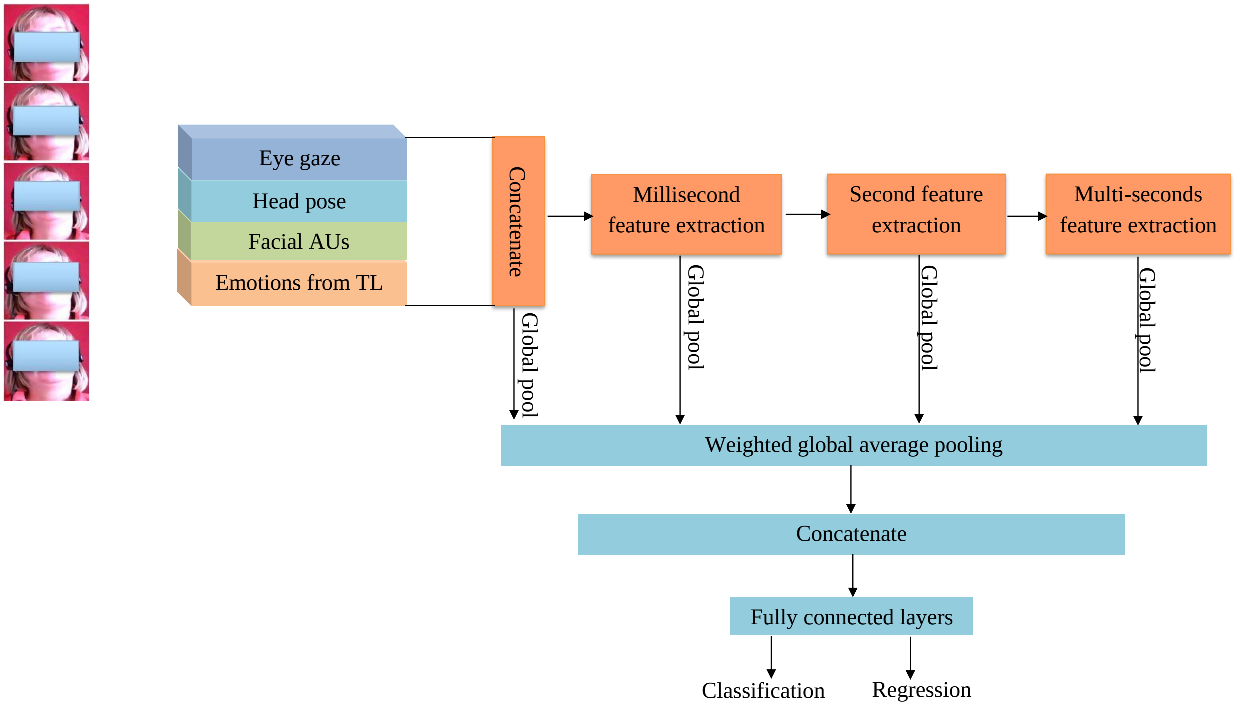


Figure 6.1 A general overview of the proposed modelling approach.

6.2 Deep learning architecture overview

Facial micro-expressions occur at short time scales of up to 500 milliseconds, and subtle intensity [273]. In mental disorders, these micro-expressions can be regraded as temporal events which can happen at different scales. This micro-expression carry useful information about behavioural patterns in these disorders. In the AVEC2018 dataset these micro-expressions are contained in short sequences of the available visual features such as eye gaze, head pose and facial action units (AUs), that can be captured using deep learning networks. The generic architecture, shown in Figure 6.1, represents the proposed approach for hierarchical extraction and interpretation of indicative time-domain sequences at multiple time scales. This concept can be applied to other mental conditions and data modalities and obtain a feature hierarchy directly from data without the domain knowledge about the related symptoms. The following sections overview the proposed network output structure and the investigated types of temporal feature extraction modules.

6.3 The proposed network output structure

The proposed network output consists of a stack of regularised fully connected layers that support dimensionality reduction, i.e. feature selection and fusion, and the calculation of regression and ternary classification output. The following sections describe several network output structure variations, which aim to incorporate the knowledge about how the observations were labelled and the outcomes of analysis discussed in chapter 5.

6.3.1 Hand-crafted features

Feature engineering is the most crucial step required to make initial input more amenable to processing machine learning algorithms. This step requires manually engineer good representations of data. On the other hand, deep learning completely automates feature engineering replacing sophisticated multistage machine learning pipelines with end-to-end models. These models can automatically learn filters at different levels that can come up with its own mechanism for feature extraction. Hand-crafted features are designed to capture descriptive characteristics of a given subject e.g.image colour or texture. Given a large number of network parameters, Deep Neural Network (DNN) optimisation space can be huge. Building on this concept, it is considered that hand-crafted features can supply complementary information for deep features. To investigate this concept, pre-summarised features from experiments in chapter 5 (section 5.6.3) were concatenated with output from the developed DNN model. These features were selected based on the highest Unweighted Average Recall (UAR) achieved using 9 minute time windows.

6.3.2 Residual learning

Classification and regression predictions obtained from experiments in section 5.6.3 were utilised in the output structure of proposed DNN architecture as given in equation 6.3.1 below :

$$\hat{Y} = \hat{Y}_{ET} + \Delta Y \tag{6.3.1}$$

Where $\hat{Y}_{ET}$ is the label predicted by an Extra Random Trees estimator described in section (5.6) and ΔY is the output of the proposed DNN. From this equation, it can be seen that the proposed DNN architecture aims to adjust the weak Extra Trees prediction. This can be considered an estimator boosting mechanism. Figure 6.2 illustrates the proposed mechanism.

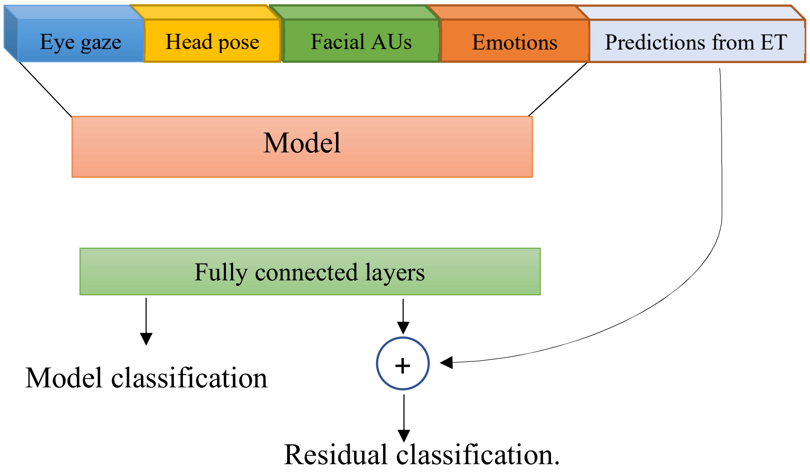


Figure 6.2 A general overview of the proposed residual learning technique.

6.3.3 Translating regression into classification

In the AVEC2018 dataset, the classes are drawn from different ranges of YMRS scores, as discussed in section 5.3. For each given class in the dataset, there are different ranges of YMRS scores that represent each of the three classes (remission, Hypo-mania and mania). However, there is ambiguity at the classes' boundary; for example, it is not clear whether YMRS score 7 is remission or hypo-mania. This means that the three classes are not completely separated based on the subjective self-reported measure. Based on this fact, the translation of regression targets into remission, hypo-mania and mania classes using the fuzzy membership functions as shown in Figure 6.3 proposed. In this way, classes will have a meaningful representation of YMRS scores. To implement this approach, two boundary values between the classes (7,20) were used. If a value is passed, an estimation is returned, which tells if the predicted YMRS value is completely related to a specific class or represents a combination of two classes. Three membership function based on sigmoid were employed. By adding residual learning to the model output and translating regression to classification, the proposed model produced six different outputs, namely: direct model classification, direct model regression, direct model residual regression to classification, residual classification, residual regression and residual regression to classification. By having these different outputs, the feature extraction part of the proposed model is trained to be useful for all of the output and not specialised toward generating useful features for a specific type of output. In this way, the model will be able to generalise for all the outputs.

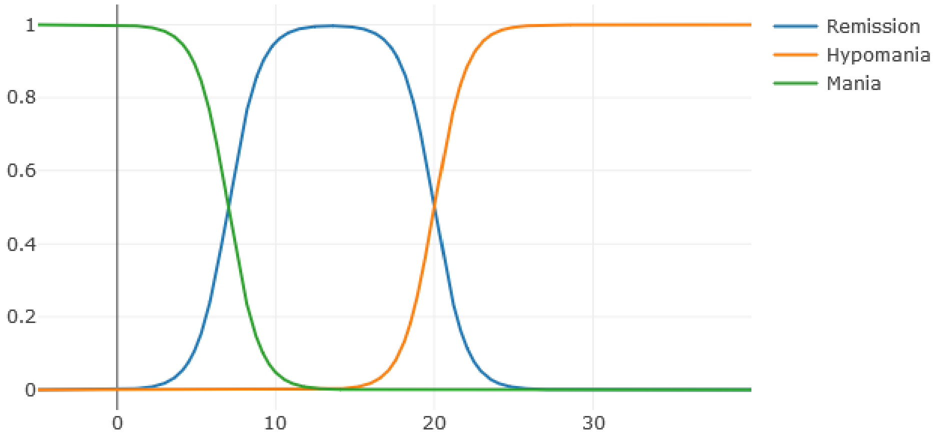


Figure 6.3 Class membership functions for regression to classification translation.

6.4 Investigated deep learning architectures

In this section, several deep learning architectures for temporal feature extraction were investigated. First, a conventional CNN model is created using a series of 1D convolutional layers (Conv1D) with maxpoling layers for downsampling. As Conv1D layers are limited to a fixed filter size, the second experiment employed multiple stacks of parallel convolutional layer blocks of different scales inspired by GoogLeNet Inception module [272]. Finally, group convolutions and modules based on ResNext [271] are evaluated.

6.4.1 CNN architecture for Sequence extraction.

First, a standard CNN architecture of stacked convolutional and downsampling layers was evaluated as shown in Figure 6.4. The aim is to develop a feature extraction sequence that can characterise useful features at different time scales according to the number of employed stacks. The notation (Conv1D, 3, 32, /1) represents a 1D convolutional layer with filter size 3, the number of filters 32 applied with strides 1, representing the time shift step.

Figure 6.5 illustrates the concept of time-scaled feature extraction approach. Extracting features at various scales using 1D CNN layers will produce a high dimensional time series data, and features from different epochs may often have different scales. Such effects will impact the fully connected layers that process CNN features and produces classification and regression outputs. This will make a direct analysis of temporal data difficult. This issue can be fixed by introducing a global average pooling operation at the link between CNN and fully connected layers representing the maximum, mean and standard deviation scope of the feature over time. This model performs combined time series analysis with conventional hand-crafted features and transfers learning model features. Output from these stacks is

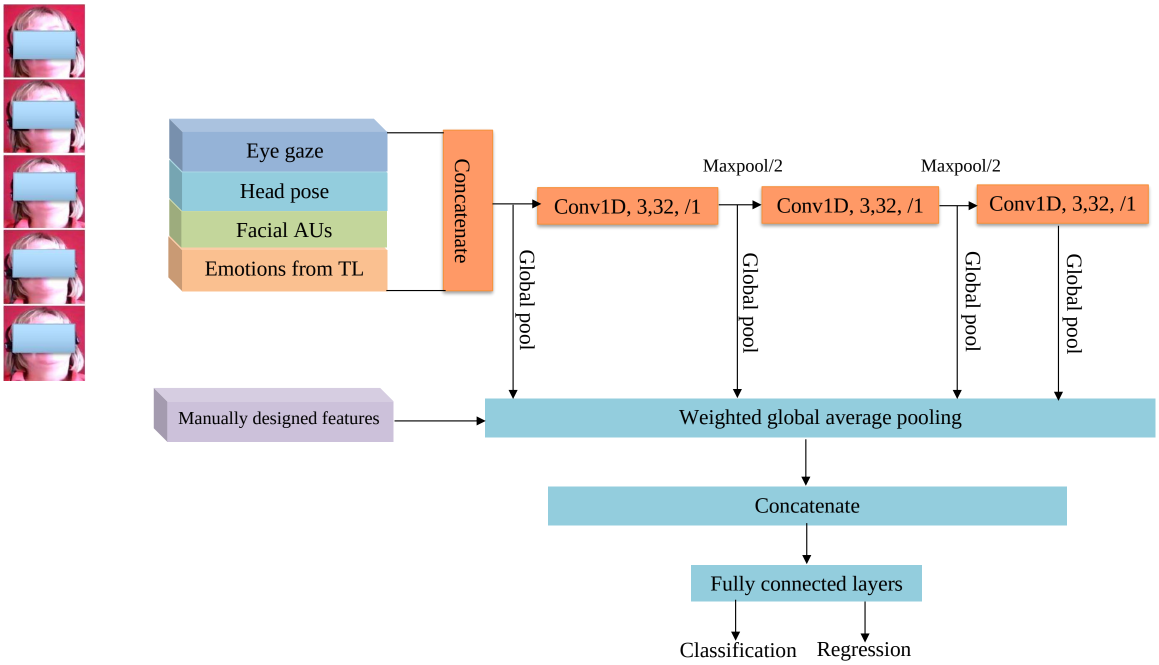


Figure 6.4 CNN architecture for Sequence extraction (CLF: Classification, Reg2clf: Regression to Classification, Reg: Regression).

concatenated and followed by batch normalisation [274] and non-linear Rectified Linear Unit [275]. Table 6.1 shows the model evaluation results on the development set for different outputs (direct and residual). Three results were obtained for each output type, namely classification (clf), regression (reg) and regression to classification (reg2clf).

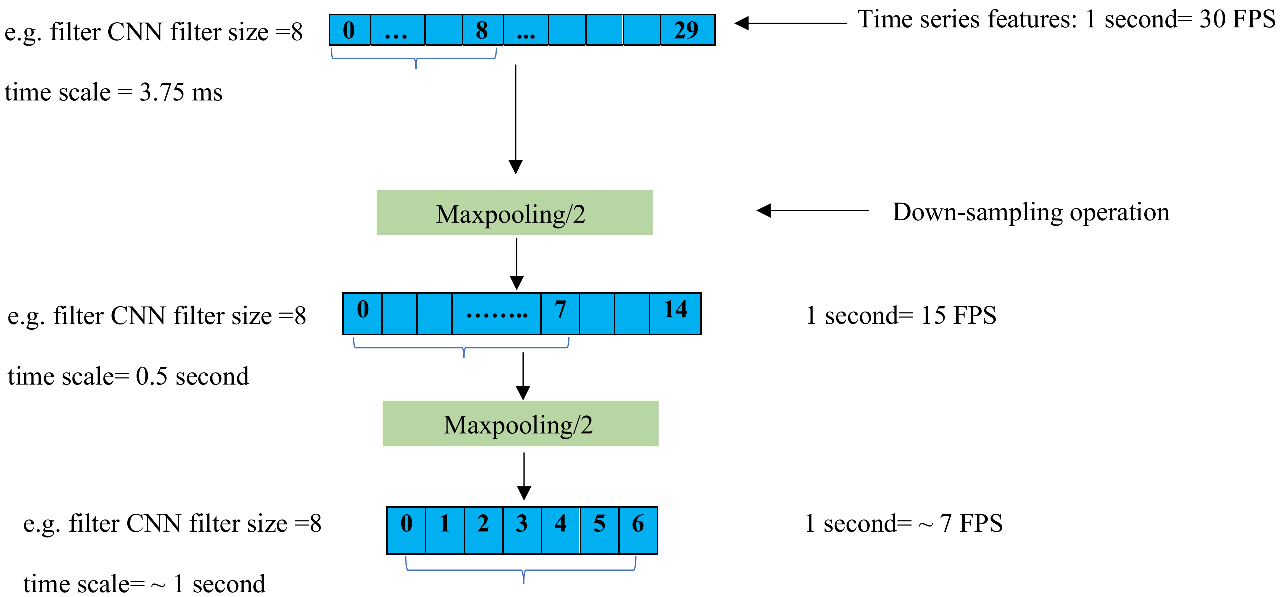


Figure 6.5 Proposed multi-time scale feature extraction.

Table 6.1 Unweighted Average Recall (UAR%) and mean absolute error from evaluation on the development set using the CNN sequence extraction model.

Classification Outputs	**Result, UAR %**
Model classification	53.2
Residual regression to classification	53.50
Residual classification	55
Regression outputs	**MAE, YMRS units**
Model regression	12.2
Residual regression	11.5

6.4.2 Inception architecture for temporal modelling

As the architecture presented in the previous section used a single Conv1D layer as a building block in a sequential stack, this may discard valuable information from the video recording. Given that the episodic structure of visual events happens at different time scales, the extraction of time-scaled feature sequences using 1D convolutional (Conv1D) layers is difficult as there is the uncertainty of the correct feature extraction scale and filter size. To handle this issue, an architecture adopted from the GoogLeNet Inception module [272] for the temporal case was developed. This architecture includes several parallel Conv1D filters of different scales, as shown in Figure 6.6.

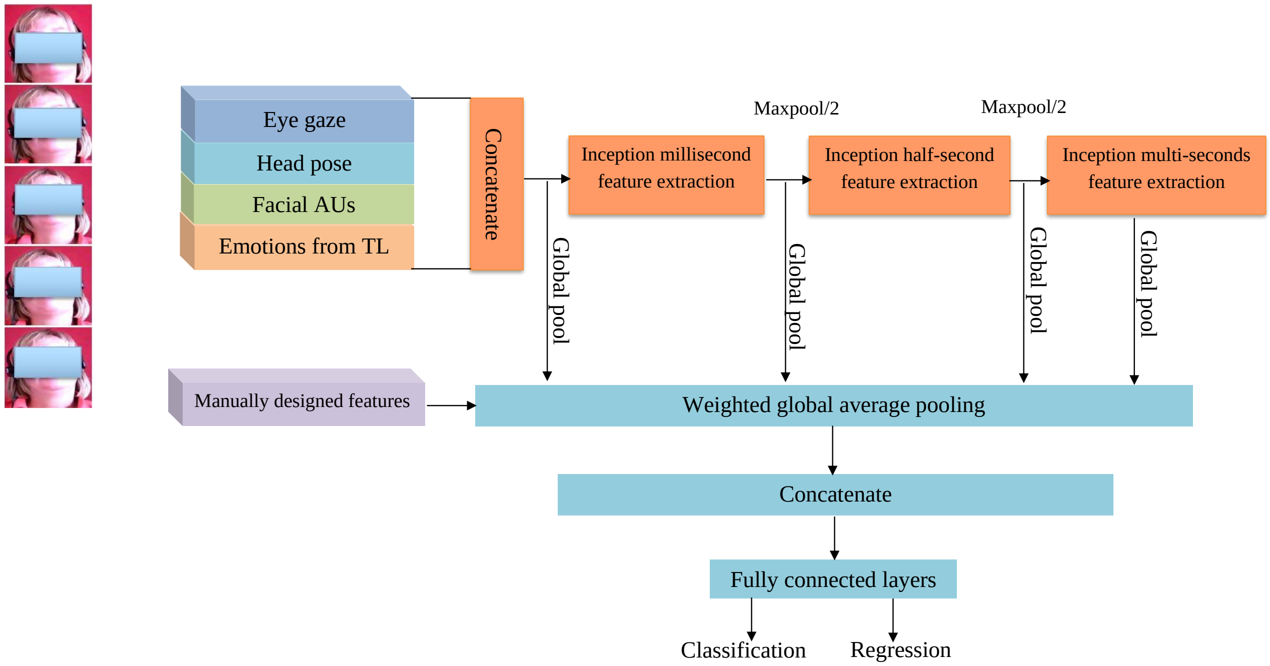

Eye gaze
Head pose
Facial AUs
Emotions from TL
Concatenate
Maxpool/2
Maxpool/2
Inception millisecond feature extraction
Inception half-second feature extraction
Inception multi-seconds feature extraction
Global pool
Global pool
Global pool
Global pool
Manually designed features
Weighted global average pooling
Concatenate
Fully connected layers
Classification
Regression

Figure 6.6 Inception module architecture.

The outputs from the parallel pathways are concatenated by the depth dimension and followed by batch normalisation [274] and non-linear Rectified Linear Unit [275]. The Inception module usage in the event extraction sequence is shown in Figure 6.7. In the proposed variant of Inception module, four parallel logarithmically scaled Conv1D filters, and a MaxPool/2 operation is used as building blocks. In each block, the linear combination of the input feature map is calculated, producing a high dimension feature vector which is downsampled by applying maxpooling. In this way, features are extracted at multi time scales to find useful representations. Next, inception module is used again to find a new representation of the downsampled data followed by batch normalisation [274] and a non-linear Rectified Unit [275]. Repeating this structure creates a set of time-scaled feature extraction pathways.

Depending on the used number of sequence extraction blocks, features can be extracted on multiple time scales. Three sequence extraction blocks have been employed to capture visual features on three-time scales, wherein each sequence extraction block features are extracted according to the Conv1D filter size. In the first sequence, block features are extracted in the range of 3.75 milliseconds given that 8 is that largest filter size. After max pooling operation, the next sequence extraction block will extract features in the range of 0.5 seconds. To control the dimensionality after the inception module, an additional Conv1D layer with ReLu activation was utilised. Table 6.2 shows the AVEC2018 development dataset UAR results obtained from the proposed temporal Inception model.

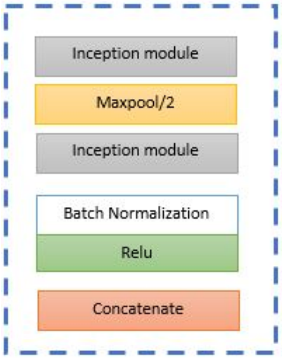


Figure 6.7 Sequence extraction block.

Table 6.2 Unweighted Average Recall (UAR%) and regression results for the development set using inception sequence extraction model.

Classification Outputs	**Result, UAR %**
Direct model classification	56%
Direct model regression to classification	58%
Residual classification	61%
Residual regression to classification	55.3%

Regression outputs	**MAE, YMRS units**
Model regression	11.3
Residual regression	8.5

6.4.3 ResNext for temporal modelling

The Residual Network (ResNet) architecture [106] have demonstrated a significant performance increase in image classification and object detection tasks. This architecture introduced a new network structure consisting of several shallow units called residual units. These units are composed of convolution, Batch Normalisation (BN) and rectified linear unit (ReLU) with a shortcut connection passed to the output of the residual units as shown in Figure 6.8. In the proposed variant of Inception module, four parallel logarithmically scaled Conv1D filters, and a MaxPool/2 operation is used as building blocks. In each block, the linear combination of the input feature map is calculated, producing a high dimension feature vector which is downsampled by applying maxpooling. In this way, features are extracted at multi time scales to find useful representations. This technique solved the gradient fading problem and resulted in the ability to train much deeper networks than what was previously possible. A ResNet with two stacked blocks is defined as:

$$y = F(x, W_i + x) \tag{6.4.1}$$

where *x* is the input to building block, y is the output of the residual block and *F(x,Wi)* is the residual mapping to be learned. This can be implemented as Feed Forward Neural Network (FFNN) with identity shortcut output added to the output of the stacked layers.

Inspired by this concept, [271] proposed a deep neural network called ResNext. This architecture has a combination of stacked blocks of ResNet and the split-transform-merge strategy behind Inception modules. Grouped convolution concept from [100] was employed in ResNext. It consists of creating a deep network with some layers and then replicating it so that there are more than 1 pathways for convolutions on a single image. Then the resulting feature outputs are aggregated by summation. A new hyperparameter is also proposed called cardinality, which refers to the number of intermediate convolutional layer groups to provide a new way of adjusting the model capacity. In their experiments, Xie et al. [271] showed that increasing cardinality has been more useful in increasing accuracy rather than using a wider or deeper network. Inspired by ResNext, a novel CNN architecture for temporal sequence extraction is proposed, as shown in Figure 6.9.

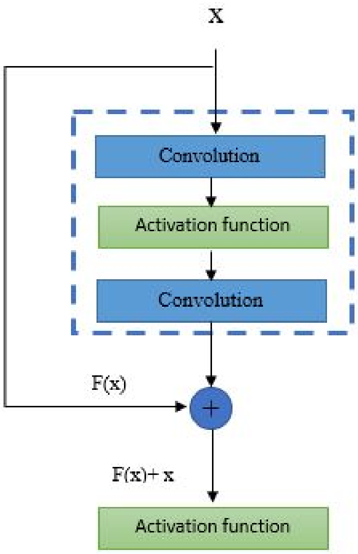


Figure 6.8 ResNet architecture proposed by [106].

This proposed architecture allows for a much deeper model to be trained while having the same amount of observation. This is useful in the case of the AVEC2018 dataset as there is a limited number of samples. The adapted architecture uses stacks of aggregated residual

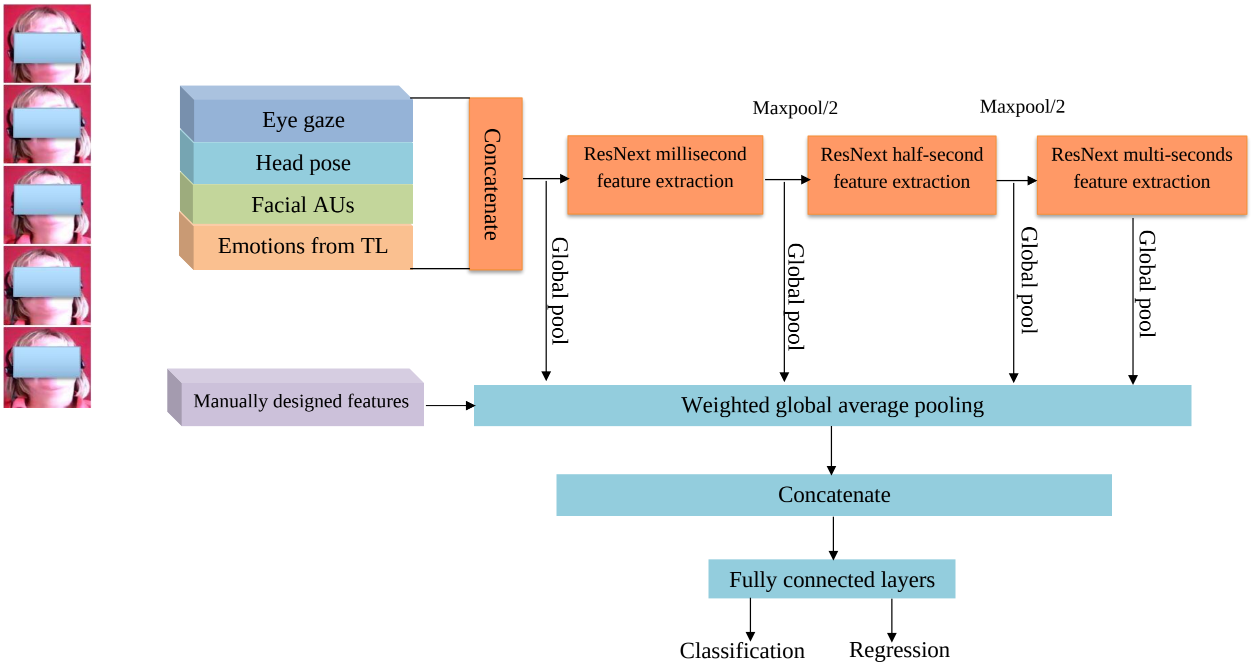

Eye gaze
Head pose
Facial AUs
Emotions from TL
Concatenate
Maxpool/2
Maxpool/2
ResNext millisecond feature extraction
ResNext half-second feature extraction
ResNext multi-seconds feature extraction
Global pool
Global pool
Global pool
Global pool
Manually designed features
Weighted global average pooling
Concatenate
Fully connected layers
Classification
Regression

Figure 6.9 Resnext architecture for Sequence extraction (CLF: Classification, Reg2clf: Regression to Classification, Reg: Regression).

transformations where each stack features are processed in different time scale. This is based on the assumption that the indicative behavioural events happen at different time scales. Residual learning skip connection was used as discussed in section 6.3.2 to enable the network to start from an optimised feature space and improve classification and regression accuracy. The first building block of the proposed architecture is the bottleneck block Figure 6.10 and Table 6.3 describes the proposed architecture.

Where the incoming time series are divided into low dimensional embedding. A set of transformations are performed on each path, and finally, the output is aggregated by summation. As the input time series data is composed of different features, it is important to pre-process them separately without mixing the channels. To implement this, the input to the proposed network is processed using depthwise separable convolutions which perform spatial convolution independently over every channel of the input time series. This is followed by a regular convolution with 1x1 windows, projecting the channels computed by the depthwise convolution onto a new channel space. After the pre-processing layer, many activation maps are generated by the network (group2) using convolutional filters (32x5x5). Groups 3-7 correspond to the residual module, where each group is a stack of residual blocks, as shown in Figure 6.9. There are three convolutional layers in each group with filter sizes 1x1 and 3x3. Batch normalisation and ReLu activation layer are applied after each residual block transformations follow the same topology as ResNext, and the size of such transformations is

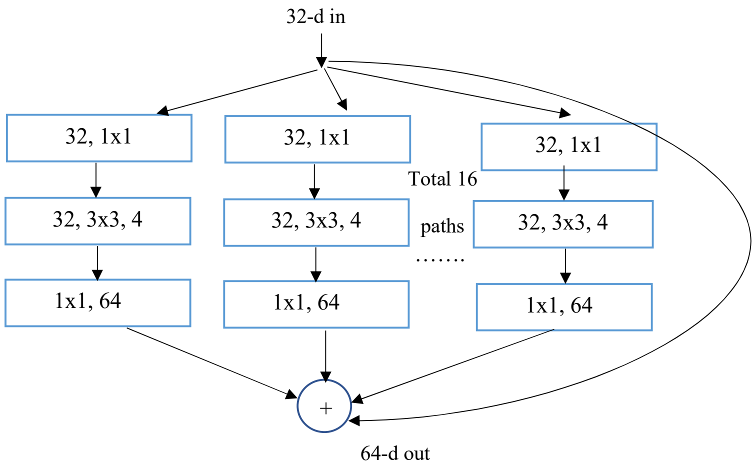


Figure 6.10 ResNext module architecture proposed in [271].

16. Global weighted average pooling is applied to the time-series features from ResNext stacks. Then, the output features from this operation are concatenated with the pre-summarised features and passed to the six different output, two of which are softmax-activated. This layer gives the probability of distribution over the three class labels. Table 6.4 shows results obtained from employing ResNext sequence extraction model on the AVEC2018 development set.

Table 6.3 Proposed ResNext architecture. Brackets represent the shape of a residual block. A layer is shown as (Filter size, Output channels).

Group	**Output size**	**Process**	**Times**
Group 1	32x32	SeparableConv1D	x1
Group 2	32x32	5x5,32, stride 2	x1
Group 3	32x64	1x1, 32 3x3, 32, C=16 1x1, 64	x 3
Group 4	32x128	1x1, 64 3x3, 64, C=16 1x1, 128	x 3
Group 5	28x256	1x1, 128 3x3, 128, C=16 1x1, 256	x 3
Group 6	256x512	1x1, 256 3x3, 256, C=16 1x1, 512	x 3
Group 7	512x1024	1x1, 512 3x3, 512, C=16 1x1, 1024	x 3
Group 7	1x1	Weighted global average pool, Fully connected layers , Softmax	x 1

Table 6.4 Unweighted Average Recall (UAR%) and regression results for the development set using ResNext sequence extraction model.

Classification outputs	**Result, UAR %**
Direct model classification	51.3%
Direct model regression to classification	61%
Residual classification	66 %
Residual regression to classification	63%
Regression outputs	**MAE, YMRS units**
Model regression	10.5
Residual regression	7.5

6.4.4 Hand-crafted feature modelling experiment

In this experiment, the capability of mapping arbitrary time series data into manually extracted statistical functionals is investigated using the three proposed architectures in previous sections. The aim is to utilise the proposed models' discriminative features to be employed as a pre-trained feature extractor model. This will allow passing time-series input through it and test the model performance using the AVEC2018 dataset.

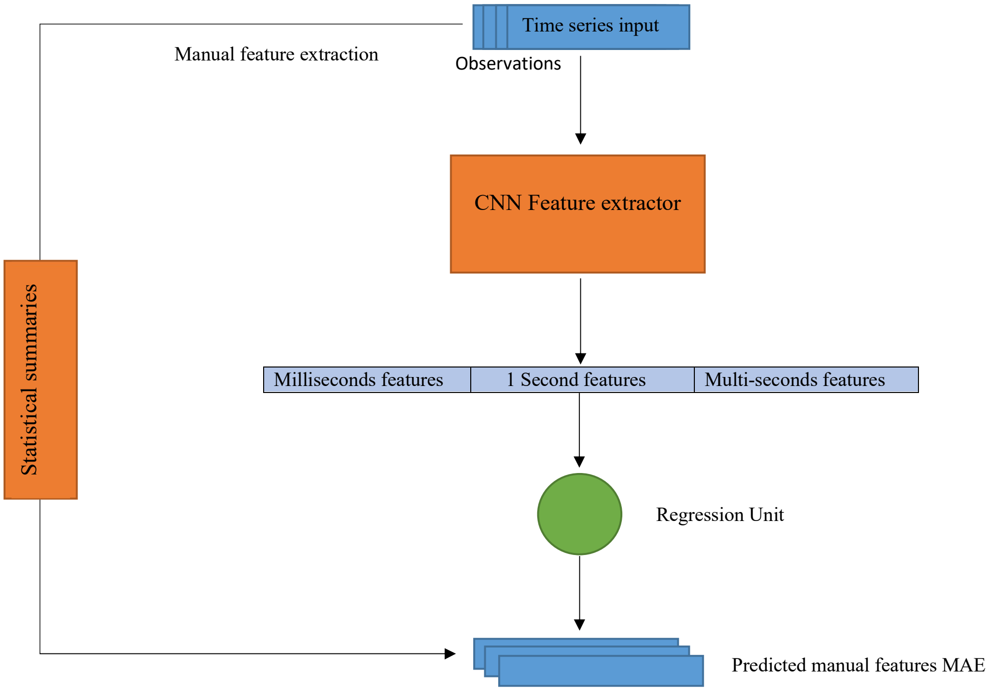


Figure 6.11 The proposed statistical functional approximation test.

As shown in Figure 6.11, the input for these thee structures is represented by robust scaled time-series observations and hand-crafted features are summarised and fed as a target for the model. Within this experiment, ResNext feature extractor (section 6.4.3) is compared with Inception (section 6.4.2) and CNN temporal modelling blocks (6.4.1). Using the multi-scale sequence extraction, the performance of these models were evaluated by MAE for each given feature statistical summary. The three network architectures' performance for predicting statistical time series summaries is shown in Table 6.5. ResNext network outperforms the network based on multiple stacks of CNNs and ResNe model. This means that ResNext can accurately model most of the target statistical summaries. Based on these results, ResNext model was used as a feature extractor model for the AVEC2018 dataset, as shown in Figure 6.12. Table 6.6 shows results obtained from this experiment on the AVEC2018 development set.

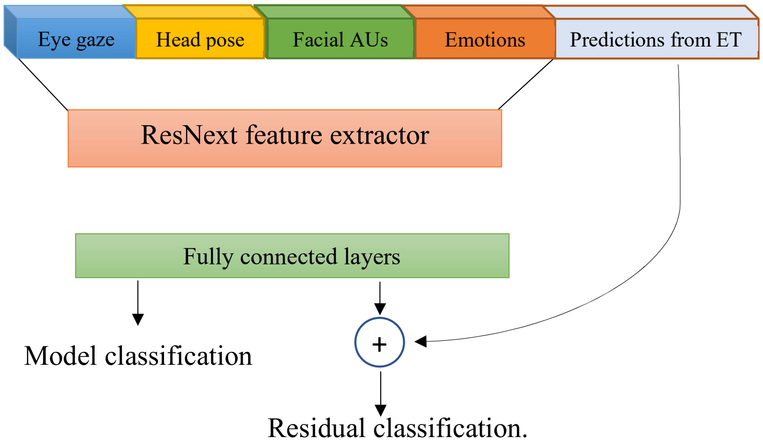


Figure 6.12 Proposed architecture for pre-trained hand-crafted feature extractor based on ResNext model (CLF: Classification, Reg2clf: Regression to Classification, Reg: Regression).

Table 6.5 Statistical feature approximation results.

Approximator	Regression MAE
CNN	0.46
Inception	0.26
ResNext	**0.16**

Table 6.6 Unweighted Average Recall (UAR%) and regression results for the development set using ResNext hand-crafted features estimator as a pre-trained feature extractor model.

Classification Outputs	**Result, UAR**
Direct model classification	35.59%
Direct model regression to classification	49.15%
Residual classification	64.40
Residual regression to classification	66.10%
Regression outputs	**MAE, YMRS units**
Model regression	10.5
Residual regression	7.5

6.5 The application of the proposed model to automated depression diagnosis

The AVEC2014 and the AVEC2013 sub-challenges aimed to detect signs of depression from patients' visual and vocal cues. Many participants proposed their methods to find a solution for this task [199] [198] [202] [276]. LGBP-TOP features [277] extracted by [209] to generate facial and Local Phase Quantisation (LPQ) features descriptor for the inner facial regions that correspond to eyes and mouth area. These features were then utilised for regression using Canonical Correlation Analysis (CCA) [278], generating an ensemble for depression score prediction. These techniques use hand-crafted visual predictors and algorithms to search for noticeable texture information, edges, and surface changes. Recently, deep learning has been employed for depression analysis. For example, features from audio information, face shape and facial appearance were investigated in [276]. For audio, YAAFE tool [279] was used to extract features such as Mel-Frequency Cepstral Coefficients (MFCC) and Linear predictive coding (LPC). For the Face appearance information, [276] pre-trained a DNN network on the FER 2013 dataset and used it to extract face appearance information. Finally, for the facial shape information, 49 facial landmarks were used to extract head movement and pose features. Temporal dynamics of these multimodal features were modelled using Long Short Term Memory (LSTM) network. Improvements in RMSE and MAE compared to the earlier approaches was reported, but the model did not outperform the state-of-the-art results of Williamson et al. [280]. The recent method proposed by Zhu et al. [208] introduced a twostream deep CNNs with joint-tuning layers for depression prediction using facial appearance features and facial dynamics features. The facial appearance features were extracted from a GoogleNet model [272] pre-trained on the public CASIA WebFace database [281]. The facial dynamics network was trained on the motion changes obtained from the optical flow images each summarising 10 consecutive video frames. Considering the time scales of indicative facial micro-expressions, this method may not be adequate to represent the subtle emotional changes in depression. The following experiment intends to test the generalisability of the proposed temporal feature extraction architectures (Conv1D, ResNext and Inception) using AVEC 2014 depression dataset which includes 300 video recordings of 83 individuals continuously assessed against individual Beck Depression Inventory-II (BDI-II) scores [29].

As shown in Figure 6.13, DNN model refers to the architecture presented in section (6.4.3). In this architecture, features were extracted at different time scales to observe the temporal events that occurs dynamically along the feature sequence. Results for each of the proposed architectures are presented in Table 6.7. Using ResNext architecture (6.4.3), the RMSE was 8.5 on test set while the MAE was 7.65 on test set. These results demonstrate the effectiveness of the both models and especially ResNext-based one which achieved performance comparable to the state-of-the-art. Results achieved on the AVEC2014 test set enables the comparison of the proposed architectures with the previously reported results, listed in Table 6.8. Results indicate that the proposed architecture outperforms most of the published methods and delivers state-of-the-art performance using video data only without audio cues. However, in the shown results audio features are utilised in many of them. This experiment indicates that processing temporal information is important for depression diagnosis and ResNext model characterizes the visual features dynamics efficiently.

Table 6.7 Continuous depression assessment result on the AVEC2014 using Conv1D, GoogleNet inception [272] and ResNext [271].

Model	**MAE**
Conv1D	11.5
GoogleNet inception	8.5
ResNext	7.65

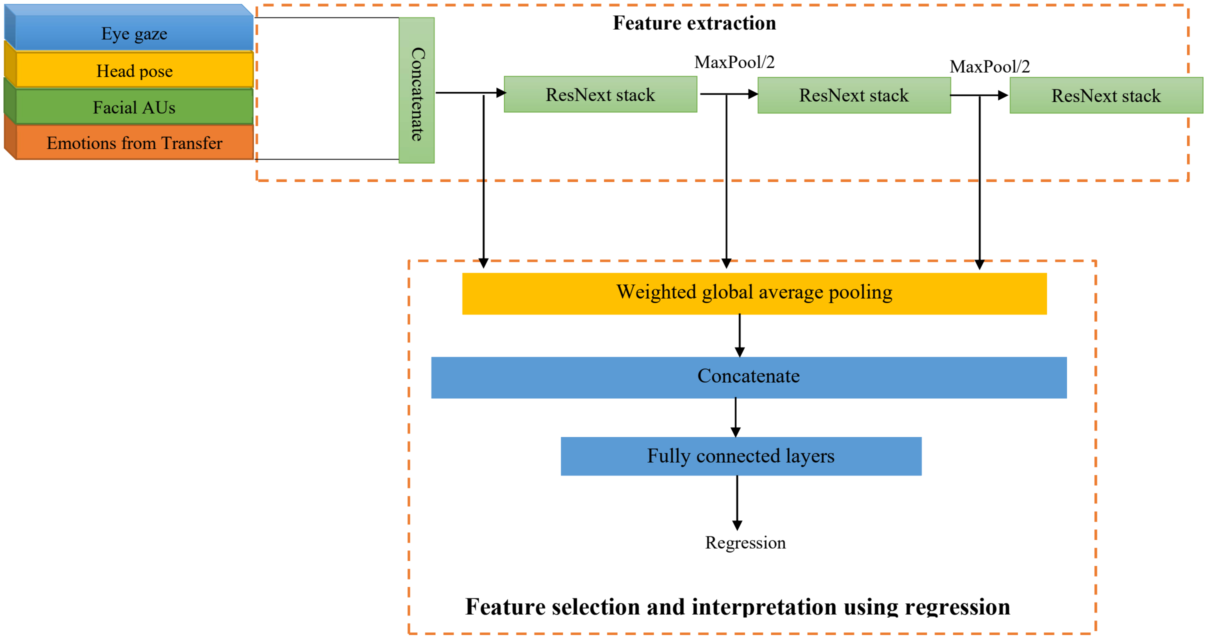


Figure 6.13 The proposed architecture for the AVEC2014 depression diagnosis sub-challenge.

Table 6.8 Accuracy comparison of the published BDI-II regression approaches obtained from evaluation on the AVEC2014 test set.

Methods	RMSE	MAE
Baseline [168]	10.86	8.86
UUIMSidorov [282]	13.87	11.20
InaoeBuap [200]	11.91	9.35
Brunel [199]	10.5	8.44
BU-CMPE [209]	9.97	7.96
Kaya [209]	10.2	8.20
zhu [208]	9.55	7.47
Proposed method	9.75	7.65

6.5.1 Transfer learning with the AVEC2014 dataset

To utilise the results achieved on the AVEC2014 dataset, transfer learning technique to obtain depression features from bipolar disorder data by passing it through the AVEC2014 depression-trained model. To this end, input features are passed through the AVEC2014 model and new features are generated from the layer before the fully connected layer in the AVEC2014 ResNext-based model. These features provide a useful interpretation of bipolar disorder data from the depression diagnosis point of view. This is inspired by the DSM-5 manual [15] which states that bipolar disorder is identified by recurring periods of depression followed by mania incidents. Similarly to the time series summaries, the extracted transfer learning features are used as an additional input to the ResNext-adapted model concatenation layer, as shown in Figure 6.14 below. Using the ResNext-adapted model, features were extracted at multi time scales. First, features were extracted at the range of approximately 15 seconds using six stacks of ResNext-model. Then three stakes are used to extract features at the range of 1 second. As shown in Table 6.9 and Table 6.10, fewer stacks produced the best result for remission, hypo-mania, mania classification, and regression. This might highlight the fact that useful events in mental disorders happen in a concise period of time. Additionally, results confirm that there is a relationship between depression symptoms and bipolar disorders.

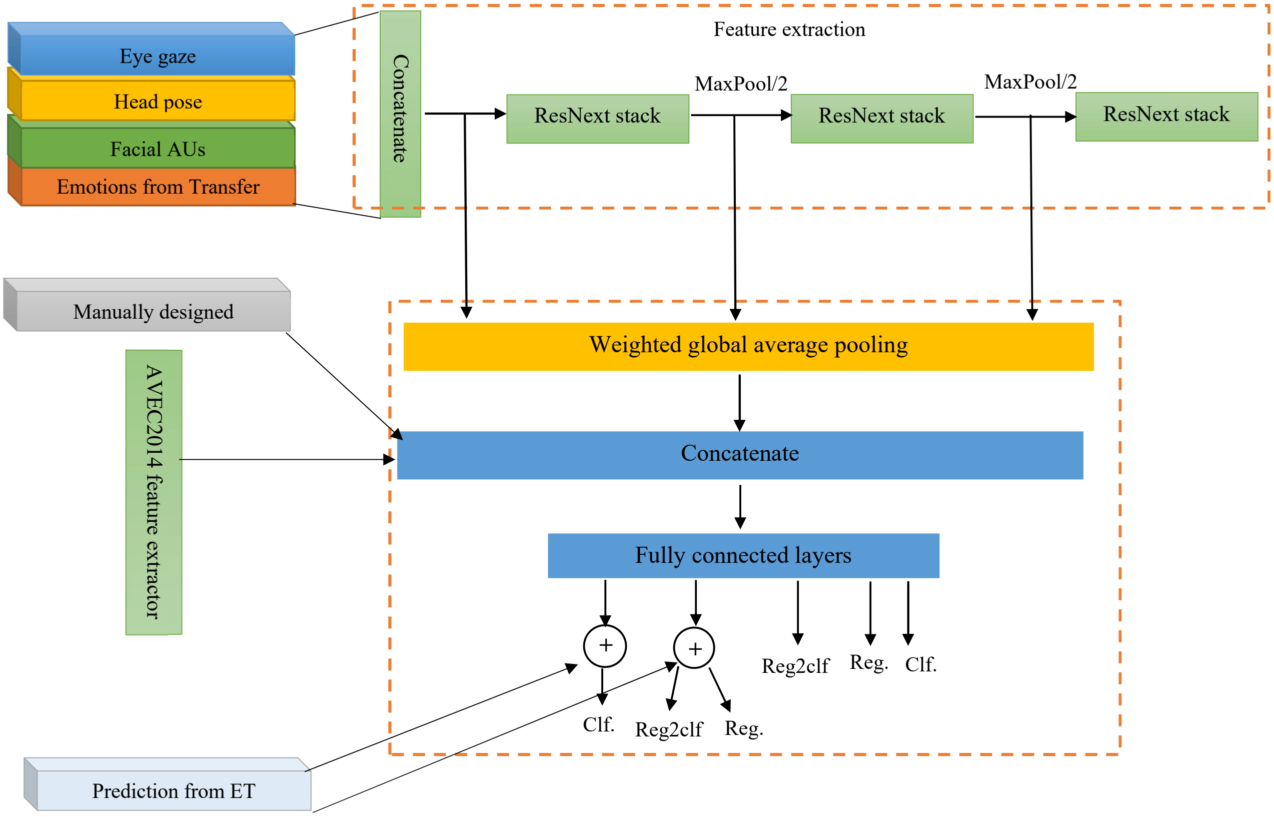


Figure 6.14 Adding transfer learning features from AVEC2014 model into ResNext Bipolar model. (CLF: Classification, Reg2clf: Regression to Classification, Reg: Regression).

Table 6.9 Unweighted Average Recall (UAR%) and regression (MAE) results for the AVEC 2018 development set using feature extracted at the range of 1 second by ResNext temporal model utilising from features extracted by the pre-trained model on AVEC 2014.

Outputs	**Conv1D**	**Inception**	**ResNext**
Direct model classification	45.56%	48.12%	50.84%
Direct model regression to classification	50.33%	53.20%	61.33%
Model regression	11.5	10.33	7.6
Residual classification	52.40%	61.2%	**69.50%**
Residual regression to classification	50.10%	60.33%	**66.10%**
Residual regression	10.8	9.5	**6.8**

Table 6.10 Unweighted Average Recall (UAR%) and regression (MAE) results for the the AVEC2018 development set using feature extracted at the range of 15 seconds by ResNext temporal model utilising from features extracted by the pre-trained model on the AVEC2014.

Outputs	**Conv1D**	**Inception**	**ResNext**
Direct model classification	42.44%	46.30%	51.65%
Direct model regression to classification	49.65%	51.33%	60.10%
Model regression	12.1	10.66	8.4
Residual classification	51.33%	60.20%	62.53%
Residual regression to classification	49.15%	59.31%	61.5
Residual regression	11.5	9.8	7.4

6.5.2 Using Audio/Video modalities for bipolar disorder diagnosis

In this experiment, audio features from Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) were utilised using ResNext temporal model (section 6.4.3) to build a temporal audio model. Using the best model, features extracted the last fully connected layers. These features were concatenated with the weighted global average pooling of the temporal ResNext model for visual features as shown in Figure 6.15 below. This model architecture has achieved the best UAR results on the AVEC2018 development set, as shown in Table 6.11 below.

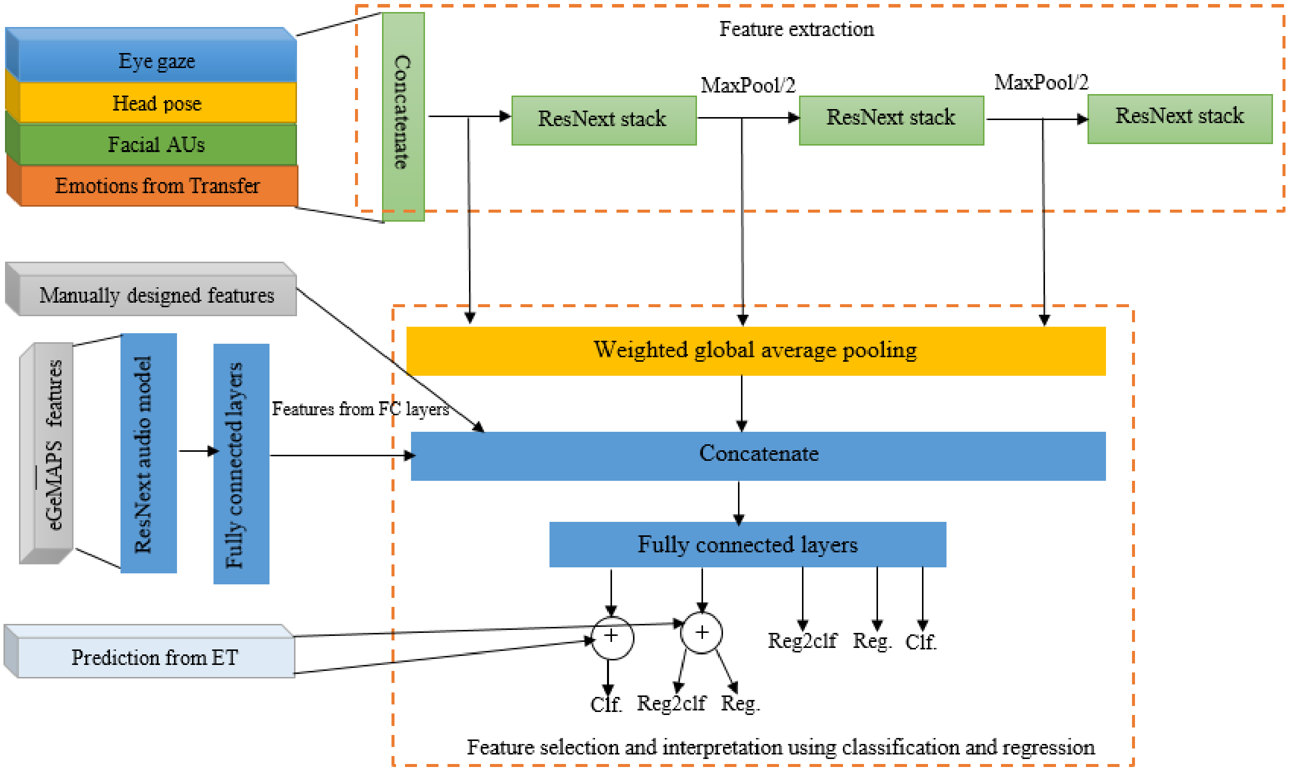


Figure 6.15 Adding transfer learning features from AVEC2018 audio model into AVEC 2018 visual feature model. (CLF: Classification, Reg2clf: Regression to Classification, Reg: Regression).

Table 6.11 Unweighted Average Recall (UAR%) and regression results for the development set using ResNext sequence extraction model with audio features from transfer learning model.

Classification outputs	**Result, UAR %**
Direct model classification	60.3%
Direct model regression to classification	63%
Residual classification	**74** %
Residual regression to classification	67%
Regression outputs	**MAE, YMRS units**
Model regression	10.3
Residual regression	6.70

6.6 Results

Several experiments were performed to find out the effectiveness of the proposed deep learning architectures in previous sections. Different time scales for feature extraction have been explored in the proposed deep learning architectures. Extraction of features on a scale of approximately 1 second provided higher classification and regression accuracies for the residual learning outputs. In contrast, the direct model classification and regression output branches did not provide better results in any considered time windows. This might indicate that important visual events happen at a very subtle scale (i.e., 1/25s to 1/5s,40 milliseconds), which is very hard to be noticed by clinicians. ResNext-based architecture with transfer learning features from eGeMAPS model achieved the highest classification (UAR) and regression (MAE) results as shown in Table Table 6.12 below. This model was utilised to predict the test set labels of the AVEC2018. Results showed comparable performance to the state-of-the-art baseline model.

Table 6.12 Comparison of bipolar classification results (UAR %) to other methods on the AVEC2018 development dataset.

Methods	**Developement UAR %**	**Test UAR%**
Baseline video[120]	55.82%	46.30%
Baseline audio[120]	55.03%	50%
Baseline video + audio[120]	60.32%	57.4%
Syed et al. [261]	N/A	57.41%
Du et al. [121]	65.51 %	57.41%
Yang et al. [260]	71.41 %	57.41%
Proposed method	64%	55.56%

6.7 Discussion

Experiments in this chapter aimed to develop an automated BD screening tool using video features. The test partition labels kept by the organisers of the AVEC2018 workshop to enable participants to develop methods for automated screening using training and development partitions and test the performance of their developed models on the test partition. As shown in Table 6.12, the test set's visual modality results outperformed the baseline results on development and test datasets for both modalities. Although the test set results for both modalities were lower than the baseline by a small margin, the experiments show that the proposed model is generalised on both depression and BD datasets. Therefore, this model can be used as universal features extractor for mental disorders. Comparing results with other studies, Du et al. [121] used IncepLSTM which integrates the capabilities of LTSM and a CNN based Inception module to capture temporal features from acoustic Low-Level Descriptors (LLDs) for the task of detecting the three states of BD. To improve the performance of IncepLSTM, they proposed the severity-sensitivity loss to optimise the cost of reducing the distances between the examples within the same class and increasing the distances between different classes. Du et al. compared the performance of their proposed method using three SVM models: (a) SVM classifier developed using MFCCs together with its velocity and acceleration features, (b) SVM classifier developed using eGeMAPS features [283], and (c) SVM classifier developed using DeepSpectum variables [284]. The Reported results show that InceptLSTM model achieved higher UAR than the three SVM models on the development partition. However, the performance of Du et al. model is not reported on the test partition. In another study, Xing et al. [285] proposed a method based on Gone et al. [286] who found that features from video, audio and text are effective for the task of automated depression screening. Xing et al. utilised Google Cloud Platform (GCP) to transcribe the AVEC2018 recordings as organisers of the AVEC2018 challenges did not provide interview transcripts. Next, audio/visual recordings sorted into three sets: negative, positive and natural according to the language's valence. Xing et al. computed many audio features, including Mel Frequency Cepstral Coefficients (MFCC) and eGeMAPS. Motion History Histogram feature extracted based on action units and Ekman's seven basic emotions for video modality. Analysis of Variance (ANOVA) [287] was used for feature selection since Xing et al. obtained a large set of features. Extreme Gradient Boosting (XGBoost) algorithm [288] was used to evaluate the proposed method achieving an impressive UAR of 86.77% on the development partition. However, the accuracy decreased to 57.41% on the test partition. Clearly, this sharp decrease suggests an overfitting issue of their model on the development partition. This means that the model built using a set of features that maximises the performance with development partition. Therefore, the model's performance on the test partition dropped as the features differed between the two partitions. Another study by Yang et al. [260] proposed histogram-based arousal features for the task of automatic bipolar severity estimation. However, since arousal points were not provided in the AVEC2018 dataset, the authors trained LSTM-RNN model to predict arousal scores of subjects in the AVEC2018 dataset using the AVEC2015 dataset. These scores were concatenated using histograms to create a unified global representation of all samples' arousal scores. The authors used OpenSmile toolkit [262] to compute several audio features, OpenPose [289], to capture body movements and action units. Due to the huge number of extracted features, Yang et al. used brute force forward search algorithm [290], SVM classifier and correlation-based feature selection procedure [291] to optimise the dimensionality of their features. Finally, the evaluation model built by fusing DNN and Random Forest classifier. They achieved UAR of 71.41% on the development partition and dropped to 57.41% for the test partition. Although Yang et al. used several modalities to develop several classifiers, they only achieved similar results to the baseline. Lastly, Syed et al. [261] proposed features known as "turbulence-features" to capture the facial features using OpenFace toolkit [84] to get facial movement changes and emotional variations that might happen due to the severity of bipolar states. Lastly, they introduced the Greedy Ensembles of Weighted Extreme Learning Machines (GEWELMs) as an evaluation model. Although Syed et al. model is considered complex, the best UAR achieved was 57.41% on the test partition. Additionally, the UAR on the development partition was not reported to determine whether their model has an overfitting issue.

6.8 Conclusion

This chapter was dedicated to developing bipolar disorder and depression automatic diagnosis systems based on deep learning. The proposed approach aims to model physiological predictors' complex temporal trajectories using various convolutional neural network architectures. These sequential features represent indicative behavioural patterns that occur at multiple time scales. Inspired by ResNext [271] and GoogLeNet [272] architectures, a novel multi-time scale feature extractor for audiovisual behavioural time series was developed and evaluated. Additionally, the feature extractor model performance was boosted by introducing residual learning concept, where summarised features and predictions from classical machine learning algorithms are fed into the deep learning model.

This enables deep learning model to learn supplementary features in addition to the manually extracted predictors, known from the domain research. Also, as regression and classification targets in the AVEC2018 are based on the scoring of YMRS, continuous regression targets were translated into classes of bipolar subjects. These two methods have shown a better performance than the conventional direct network output. To test whether the proposed model can generalise to other mental disorder diagnoses, the AVEC2014 depression dataset was tested. The sequence extraction architecture based on ResNext model delivered comparable performance to the state-of-the-art methods. The same architecture trained on the AVEC2014 dataset utilised as a transfer learning model to extract descriptive features for bipolar disorder as both disorders are related. This experiment achieved highest predicting for the severity of mania prediction on development set and comparable performance to the state-of-the-art on the AVEC2018 test set.

Chapter 7

Conclusions and Future Works

This thesis investigated the development of several methods for automatic mental disorder diagnosis. For this purpose, a novel eye blink detection technique had been developed to address the limitation of using eye blink features in mental disorder diagnosis. Later, this technique was used to diagnose depression disorder and combined with other visual features to discriminate neurodegenerative diseases from other patients with a functional memory disorder. Furthermore, visual-based screening was developed to model visual behaviours during a speech in BD disorders, which helped diagnose BD disorder symptoms from other health controls. Additionally, this method showed it is useful in diagnosing depression patients, which means the developed system generalises well for different mental disorders.

Recent studies have shown that mental disorders are serious health problems that impact many individuals worldwide. The World Health Organisation (WHO) considers depression as the main causes of disability worldwide, whereby an estimated 300 million people are affected by depression [3]. The same issue is applied to BD, which impacts around 60 million people worldwide [3]. Dementia is affecting more than 850,000 people in the UK with care cost exceeding £26.3 billion per year [292].

Depression is a psychiatric mental disorder result from a sudden stressful event affecting an individual's life. Depression causes a continues feeling of sadness, hard to handle everyday responsibilities and negativity. Usually, triggers suicidal thoughts to end one's life [293]. The neurodegenerative disorder is composed of a group of symptoms such as a decline in memory, judgement, reasoning and daily life activities. Also, neurodegenerative disorder affects visual appearance and speech performance. With the disease progression, patients will have difficulties in expressing their needs and commutation. This will make the patients isolate themselves from surrounding people, and others will show aggressive behaviours and develop depression.

In regards to depression diagnosis, test are based on interviews assessment and patients self-report, for example, the Hamilton Rating Scale for Depression. These tests measure the severity of symptoms and behaviours based on a test score range. However, using this method for assessment depends on the patient's ability to report their symptoms honestly. As a result, the diagnosis process is time-consuming and involves a huge amount of clinical training and trials to produce satisfactory results. Similarly, BD diagnosis depends on a questionnaire style interview to assess mania's existence and its severity. Neurodegenerative disorder diagnosis is challenging due to the memory concerns that overlap with other disorders, such as depression, functional memory disorders, or normal ageing. The tools used to diagnose patients with a high risk of developing neurodegenerative disorders are either invasive, for example, the CerebroSpinal Fluid (CSF) or costly, for example, Positron Emission Tomography (PET). Therefore, developing non-invasive, automatic, and objective screening tools that can be used frequently, easily administrated, and remotely applicable is high demand by health care providers. This tool can speed up the diagnosis of mental disorders and provide the right medication and care. This thesis has used the latest visual signal processing and machine learning algorithms to develop an automated method for screening mental disorders. To this end, the thesis attempted to find answers to research questions listed in chapter 1.

7.1 Conclusion of results

7.1.1 The feasibility of using facial landmark detectors for the development of eye blink detection algorithm

Using eye blinks for mental disorder diagnoses such as depression and memory disorders has been established in the literature. However, the techniques used in these studies were either time-consuming and not suited for clinical applications [19, 127], or invasive [230].

Therefore, to address this limitation, the first research question was how feasible it is to develop eye blink detection algorithms using facial landmarks detectors. For this purpose, Chapter 3 described a system that uses Zface facial landmark detector to localise eye region landmarks to find the distance between eyelids when eyes are open or closed. Then, the obtained signal is filtered using the Savitzky–Golay (SG) filter [164] and evaluated for blink detection. This method's effectiveness for eye blink detection has been investigated using five publicly available datasets achieving state-of-the-art results. The finding of this work was published in [294, 295] and contributed to answering the first and the second research question.

7.1.2 The feasibility of using eye features to identify depression automatically

Chapter 4 tested the efficacy of the proposed eye blink method to develop an automated method for depression diagnosis using publicly provided datasets provided by the AVEC2014 challenge on depression severity prediction [168]. Using only eye blink detection features the proposed depression evaluation method outperformed the baseline challenge results with MAE of 7.4 and RMSE of 9.10 on the development set and MAE of 8.30 and RMSE of 10.50 for the test set. The baseline results were MAE of 7.57 and RMSE of 9.31 on the development set and MAE of 8.86 and RMSE of 10.86 for the test set. Based on these findings, the extracted eye blink features showed their ability to diagnose the severity of depression with better results than the baseline video and audio modalities for the AVEC2014. Results in this chapter provide the answer for the second question, and The finding of this work was published in [296]

7.1.3 The feasibility of developing an automated system using visual features for the detection of neurodegenerative disorders (ND), mild cognitive impairment (MCI) and functional memory disorders (FMD)

The second part of chapter 5, the combination of eye blink features with visual features extracted from head movements and eye gaze, was investigated to develop ND, MCI and FMD automated diagnosis. This research developed with collaboration from neurologists at the Royal Hallamshire hospital in Sheffield. Visual features were extracted for each question during the clinical interview to assess the contribution of each question. Then, features extracted from all question were combined to investigate further how the results will change. This study's findings demonstrated that visual features could support patients' initial diagnosis with concerns related to ND. , FMD (i.e., subjective memory concerns unassociated with objective cognitive deficits or risk of progression) and Mild Cognitive Impairment (MCI). This method showed promising results with an accuracy ranged between 68% and 96% using the Adaboost algorithm. Furthermore, findings showed that visual features from question number 7 have the most distinctive features. This question is related to long-term memory, highlighting the importance of further investigating this type of questions. The results in this chapter provide the answer to the third research question.

7.1.4 The feasibility of employing behavioural micro-expression for BD?, What is the importance of extracting visual features on different time scales for automatic diagnosis of BD? and which deep learning network architecture has the potential of extracting useful behavioural patterns

Chapter 5 represents the first part of the proposed automated diagnosis for BD using the AVEC2018 bipolar prediction challenge dataset [120]. Three data partition were provided as audio/video recordings formed of 104 for training, 60 for development and 54 for the test set. Training and development recording were labelled into three classes using YMRS: remission, hypo-mania and mania. However, the labels for the test were not provided to the public. The purpose of this challenge is to develop BD screening tool to help the clinicians in diagnosis. Audio/visual features were extracted (i.e. eye gaze, head pose, AUs and facial emotions) to model BD patients' complex behaviour. The feasibility of extracted feature was investigated using machine learning and different time-window lengths to support the work on a deep learning model described in chapter 6. Results obtained from chapter 6 showed a promising performance of extracted features. In chapter 6, several convolutional neural network architectures were developed for automatic extraction of feature temporal trajectories during the speech. Inspired by ResNext [271] and Inception models [272], a novel deep learning architecture developed which can extract useful audio/visual features based on multi-scale temporal modelling to characterise the facial micro-expressions in BD. The intention was to build a multi-purpose deep learning feature extractor model to learn features from different kinds of mood disorders. This is confirmed on the experiments conducted on the AVEC2014 depression disorder dataset, which produced comparable results to the state-of-the-art results with MAE=7.65 and RMSE=9.75. Additionally, the AVEC 2014 model has been utilised as a transfer learning model to extract depression features from bipolar disorder. This experiment achieved the highest UAR classification accuracy of 74%, which outperformed baseline audio and video modalities reported in the AVEC2018 development set. This model used to predict the test set labels, and the result from the dataset owners was UAR 55.56%. This result shows the developed deep learning architecture's generalisation performance in capturing complex facial micro-expression in depression and the three BD severity levels. The work in chapter 5 and chapter 6 answered the fourth research question.

7.2 Future research

The methods described in this thesis can be considered as a great contribution to continue in the direction of automatic screening methods development based on audio/visual modalities. Results achieved demonstrate the feasibility of using automatic behaviour analysis to diagnose mood disorder conditions such as depression, bipolar and memory disorders. Below is some suggested direction to expand the work done in this thesis.

7.2.1 Eye blink detection

The proposed method for eye blink detection is based on the offline processing of eye landmarks. It would be interesting to explore the development of an online method for eye signal processing and investigate other combinations of signal filtering. Such addition might provide a useful communication tool for another type of disabilities where the eye is the only mean to communicate.

7.2.2 Automated diagnosis of Depression

In this thesis, the proposed automated diagnosis method for depression disorder employed visual modality features (i.e. eye blink features). Future work can include evaluating this approach on larger datasets for other mental disorders like anxiety, Attention deficit hyperactivity disorder (ADHD), and Autism spectrum disorder (ASD). Moreover, given the recent improvements in natural language processing (NLP) [297] [298] this feature can have an important role to investigate in the field of automated diagnosis of mental disorders. Also, features from behaviour signals like body movements have not been utilised in this thesis. Adding this modality can result in comprehensive behaviour analysis as compared to eye blink features alone.

7.2.3 Automated Screening of Bipolar

- In addition to the employed audio/visual features for automated bipolar diagnosis, the efficacy of body movement analysis can be investigated using the developed deep learning architecture for multi-scale temporal analysis. This modality agrees with the YMRS description of bipolar disorder as it shows how an individual reacts to events in normal daily life. Developing such algorithms will provide a complete behaviour analysis as compared to audio/visual alone.

- Given that there are overlapping symptoms between depression, BD and neurodegenerative disorders, the future research will explore the feasibility of using the developed deep learning architecture to diagnose depression, ND, FMD, MCI, remission, hypomania and mania. Furthermore, the system will predict the scores of PHQ, YMRS and MMSE for all the datasets. This will need to create a unified dataset from all dataset studies in this thesis, including the AVEC2014, Royall Hallamshire and the AVEC2018.

Bibliography

[1] R. Belmaker and G. Agam, “Major depressive disorder,” *New England Journal of Medicine*, vol. 358, no. 1, pp. 55–68, 2008.

[2] E. Vieta, M. Berk, T. G. Schulze, A. F. Carvalho, T. Suppes, J. R. Calabrese, K. Gao, K. W. Miskowiak, and I. Grande, “Bipolar disorders,” *Nature reviews Disease primers*, vol. 4, no. 1, pp. 1–16, 2018.

[3] W. H. Organization *et al.*, “Depression and other common mental disorders: global health estimates,” World Health Organization, Tech. Rep., 2017.

[4] W. Depression, “Other common mental disorders: global health estimates,” *Geneva: World Health Organization*, pp. 1–24, 2017.

[5] P. R. McCrone, S. Dhanasiri, A. Patel, M. Knapp, and S. Lawton-Smith, *Paying the price: the cost of mental health care in England to 2026*. King’s Fund, 2008.

[6] C. M. Thomas and S. Morris, “Cost of depression among adults in england in 2000,” *The British Journal of Psychiatry*, vol. 183, no. 6, pp. 514–519, 2003.

[7] J.-P. Lépine and M. Briley, “The increasing burden of depression,” *Neuropsychiatric disease and treatment*, vol. 7, no. Suppl 1, p. 3, 2011.

[8] V. Manicavasagar *et al.*, “A review of depression diagnosis and management,” *InPsych: The Bulletin of the Australian Psychological Society Ltd*, vol. 34, no. 1, p. 8, 2012.

[9] I. Grande, M. Berk, B. Birmaher, and E. Vieta, “Bipolar disorder,” *The Lancet*, vol. 387, no. 10027, pp. 1561–1572, 2016.

[10] J. Angst, “Bipolar disorders in dsm-5: strengths, problems and perspectives,” *International journal of bipolar disorders*, vol. 1, no. 1, p. 12, 2013.

[11] “Mental disorders.” [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/mental-disorders

[12] S. Hodge and E. Hailey, “Second english national memory clinics audit report,” *London: Royal College of Psychiatrists*, 2015.

[13] D. J. Blackburn, “A diagnosis for£ 55: what is the cost of government initiatives in dementia case finding?” *Age and Ageing*, vol. 43, no. eLetters Supplement, 2014.

[14] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri, “Comparison of beck depression inventories-ia and-ii in psychiatric outpatients,” *Journal of personality assessment*, vol. 67, no. 3, pp. 588–597, 1996.

[15] A. P. Association *et al.*, “Diagnostic and statistical manual of mental disorders,” *BMC Med*, vol. 17, pp. 133–137, 2013.

[16] R. C. Young, J. T. Biggs, V. E. Ziegler, and D. A. Meyer, “A rating scale for mania: reliability, validity and sensitivity,” *The British Journal of Psychiatry*, vol. 133, no. 5, pp. 429–435, 1978.

[17] 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,” *Medical science monitor: international medical journal of experimental and clinical research*, vol. 21, p. 646, 2015.

[18] W. J. Katon, “Clinical and health services relationships between major depression, depressive symptoms, and general medical illness,” *Biological psychiatry*, vol. 54, no. 3, pp. 216–226, 2003.

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

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

[21] H. Kaya and A. A. Salah, "Eyes whisper depression: A cca based multimodal approach," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 961–964.

[22] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE journal of biomedical and health informatics*, vol. 22, no. 2, pp. 525–536, 2018.

[23] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The phq-8 as a measure of current depression in the general population," *Journal of affective disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[24] T. Sharp and P. J. Cowen, "5-ht and depression: is the glass half-full?" *Current opinion in pharmacology*, vol. 11, no. 1, pp. 45–51, 2011.

[25] J. Gatt, C. Nemeroff, C. Dobson-Stone, R. Paul, R. Bryant, P. Schofield, E. Gordon, A. Kemp, and L. Williams, "Interactions between bdnf val66met polymorphism and early life stress predict brain and arousal pathways to syndromal depression and anxiety," *Molecular psychiatry*, vol. 14, no. 7, pp. 681–695, 2009.

[26] M. O. Poulter, L. Du, I. C. Weaver, M. Palkovits, G. Faludi, Z. Merali, M. Szyf, and H. Anisman, "Gabaa receptor promoter hypermethylation in suicide brain: implications for the involvement of epigenetic processes," *Biological psychiatry*, vol. 64, no. 8, pp. 645–652, 2008.

[27] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and vision computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[28] J. M. Girard and J. F. Cohn, "Automated audiovisual depression analysis," *Current opinion in psychology*, vol. 4, pp. 75–79, 2015.

[29] R. A. Steer, R. Ball, W. F. Ranieri, and A. T. Beck, “Dimensions of the beck depression inventory-ii in clinically depressed outpatients,” *Journal of clinical psychology*, vol. 55, no. 1, pp. 117–128, 1999.

[30] M. A. Sayette, J. F. Cohn, J. M. Wertz, M. A. Perrott, and D. J. Parrott, “A psychometric evaluation of the facial action coding system for assessing spontaneous expression,” *Journal of Nonverbal Behavior*, vol. 25, no. 3, pp. 167–185, 2001.

[31] J. F. Cohn and F. De la Torre, “Automated face analysis for affective,” in *The Oxford handbook of affective computing*, 2014, p. 131.

[32] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall, “Automated facial expressions analysis in schizophrenia: A continuous dynamic approach,” in *International Symposium on Pervasive Computing Paradigms for Mental Health*. Springer, 2015, pp. 72–81.

[33] ——, “Facial expressions and flat affect in schizophrenia, automatic analysis from depth camera data,” in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2016, pp. 220–223.

[34] 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 *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–8.

[35] M. D. Samad, N. Diawara, J. L. Bobzien, J. W. Harrington, M. A. Witherow, and K. M. Iftekharuddin, “A feasibility study of autism behavioral markers in spontaneous facial, visual, and hand movement response data,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 353–361, 2017.

[36] T. Guha, Z. Yang, A. Ramakrishna, R. B. Grossman, D. Hedley, S. Lee, and S. S. Narayanan, “On quantifying facial expression-related atypicality of children with autism spectrum disorder,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 803–807.

[37] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition.* Ieee, 2009, pp. 248–255.

[39] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[40] C. Zhang and Z. Zhang, "Improving multiview face detection with multi-task deep convolutional neural networks," in *IEEE Winter Conference on Applications of Computer Vision.* IEEE, 2014, pp. 1036–1041.

[41] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval.* ACM, 2015, pp. 643–650.

[42] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3676–3684.

[43] R. E. Jack and P. G. Schyns, "The human face as a dynamic tool for social communication," *Current Biology*, vol. 25, no. 14, pp. R621–R634, 2015.

[44] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[45] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing.* Springer, 2013, pp. 117–124.

[46] Y.-I. Tian, T. Kanade, and J. F. Cohn, “Recognizing action units for facial expression analysis,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[47] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn, “Fera 2015-second facial expression recognition and analysis challenge,” in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 6. IEEE, 2015, pp. 1–8.

[48] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, “The first facial expression recognition and analysis challenge,” in *Face and Gesture 2011*. IEEE, 2011, pp. 921–926.

[49] B. Jiang, M. F. Valstar, and M. Pantic, “Action unit detection using sparse appearance descriptors in space-time video volumes,” in *Face and Gesture 2011*. IEEE, 2011, pp. 314–321.

[50] A. Lonare and S. V. Jain, “A survey on facial expression analysis for emotion recognition,” *International journal of advanced research in computer and communication engineering*, vol. 2, no. 12, 2013.

[51] M. Marcus, T. Yasamy, M. Van Ommeren, D. Chisholm, and S. Saxena, “Depression—a global public health concern 2012,” *World Health Organization http://www. who. int/mental_health/management/depression/who_paper_depression_wfmh_2012. pdf*, 2014.

[52] H. A. Whiteford, L. Degenhardt, J. Rehm, A. J. Baxter, A. J. Ferrari, H. E. Erskine, F. J. Charlson, R. E. Norman, A. D. Flaxman, N. Johns *et al.*, “Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010,” *The Lancet*, vol. 382, no. 9904, pp. 1575–1586, 2013.

[53] M. Reddy, “Depression: the disorder and the burden,” *Indian journal of psychological medicine*, vol. 32, no. 1, p. 1, 2010.

[54] *DSM-IV-TR*. 1000 Wilson Boulevard, Arlington, VA 22209: American Psychiatric

Association, 2000. [Online]. Available: https://dsm.psychiatryonline.org/doi/abs/10.1176/appi.books.9780890420249.dsm-iv-tr

[55] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of general internal medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[56] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 147–152.

[57] G. M. Lucas, J. Gratch, S. Scherer, J. Boberg, and G. Stratou, "Towards an affective interface for assessment of psychological distress," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 539–545.

[58] R. Plutchik, "A psychoevolutionary theory of emotions," 1982.

[59] P. Ekman, "Darwin, deception, and facial expression," *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 205–221, 2003.

[60] F. T. McAndrew, "A cross-cultural study of recognition thresholds for facial expressions of emotion," *Journal of Cross-Cultural Psychology*, vol. 17, no. 2, pp. 211–224, 1986.

[61] E. A. Haggard and K. S. Isaacs, "Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy," in *Methods of research in psychotherapy*. Springer, 1966, pp. 154–165.

[62] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

[63] S. Porter and L. Ten Brinke, "Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions," *Psychological science*, vol. 19, no. 5, pp. 508–514, 2008.

[64] M. Takalkar, M. Xu, Q. Wu, and Z. Chaczko, "A survey: facial micro-expression recognition," *Multimedia Tools and Applications*, vol. 77, no. 15, pp. 19 301–19 325, 2018.

[65] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 299–310, 2015.

[66] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 254–267, 2017.

[67] Q. Li, J. Yu, T. Kurihara, and S. Zhan, "Micro-expression analysis by fusing deep convolutional neural network and optical flow," in *2018 5th International Conference on Control, Decision and Information Technologies (CoDIT)*. IEEE, 2018, pp. 265–270.

[68] S.-J. Wang, B.-J. Li, Y.-J. Liu, W.-J. Yan, X. Ou, X. Huang, F. Xu, and X. Fu, "Micro-expression recognition with small sample size by transferring long-term convolutional neural network," *Neurocomputing*, vol. 312, pp. 251–262, 2018.

[69] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3d flow convolutional neural network," *Pattern Analysis and Applications*, Nov 2018. [Online]. Available: https://doi.org/10.1007/s10044-018-0757-5

[70] B. Allaert, I. R. Ward, I. M. Bilasco, C. Djeraba, and M. Bennamoun, "Optical flow techniques for facial expression analysis: Performance evaluation and improvements," *arXiv preprint arXiv:1904.11592*, 2019.

[71] W. Guojiang, Y. Guoliang, and F. Kechang, "Facial expression recognition based on extended optical flow constraint," in *2010 International Conference on Intelligent Computation Technology and Automation*, vol. 2. IEEE, 2010, pp. 297–300.

[72] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[73] C. Tomasi and T. K. Detection, "Tracking of point features," Tech. Rep. CMU-CS-91-132, Carnegie Mellon University, Tech. Rep., 1991.

[74] J.-Y. Bouguet *et al.*, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm," *Intel Corporation*, vol. 5, no. 1-10, p. 4, 2001.

[75] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural computing and applications*, vol. 24, no. 1, pp. 175–186, 2014.

[76] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[77] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429–2437, 2004.

[78] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856–863.

[79] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[80] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine Learning Proceedings 1992*. Elsevier, 1992, pp. 249–256.

[81] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: a comparative," *J Mach Learn Res*, vol. 10, no. 66-71, p. 13, 2009.

[82] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signals, their function, and automatic analysis: a survey," in *Proceedings of the 10th international conference on Multimodal interfaces*, 2008, pp. 61–68.

[83] C. A. Corneanu, M. O. Simón, 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 transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[84] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[85] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[86] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, “Support vector regression machines,” *Advances in neural information processing systems*, vol. 9, pp. 155–161, 1996.

[87] C. M. Bishop, *Pattern recognition and machine learning.* springer, 2006.

[88] T. K. Ho, “The random subspace method for constructing decision forests,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[89] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: a new learning scheme of feedforward neural networks,” in *2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541)*, vol. 2. IEEE, 2004, pp. 985–990.

[90] D. R. Cox, “The regression analysis of binary sequences,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, 1958.

[91] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller, “Cca based feature selection with application to continuous depression recognition from acoustic speech features,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2014, pp. 3729–3733.

[92] A. Pampouchidou, O. Simantiraki, C.-M. Vazakopoulou, C. Chatzaki, M. Pediaditis, A. Maridaki, K. Marias, P. Simos, F. Yang, F. Meriaudeau *et al.*, “Facial geometry and speech analysis for depression detection,” in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* IEEE, 2017, pp. 1433–1436.

[93] 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 *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 89–96.

[94] 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 *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 43–50.

[95] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang, "A random forest regression method with selected-text feature for depression assessment," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 61–68.

[96] 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," *depression*, vol. 1, no. 1, 2014.

[97] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[98] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, "A practical guide to support vector classification," 2003.

[99] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[100] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[101] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[102] B. Fasel, "Head-pose invariant facial expression recognition using convolutional neural networks," in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*. IEEE Computer Society, 2002, p. 529.

[103] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian conference on computer vision*. Springer, 2014, pp. 143–157.

[104] H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, and J. Kim, “Deep temporal appearance-geometry network for facial expression recognition,” *arXiv preprint arXiv:1503.01532*, 2015.

[105] A. Gudi, H. E. Tasli, T. M. Den Uyl, and A. Maroulis, “Deep learning based facs action unit occurrence and intensity estimation,” in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 6. IEEE, 2015, pp. 1–5.

[106] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[107] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[108] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91–99.

[109] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[110] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[111] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[112] D. Scherer, A. Müller, and S. Behnke, “Evaluation of pooling operations in convolutional architectures for object recognition,” in *Artificial Neural Networks–ICANN 2010*. Springer, 2010, pp. 92–101.

[113] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” in *European conference on computer vision*. Springer, 2014, pp. 346–361.

[114] M. Buckland and F. Gey, “The relationship between recall and precision,” *Journal of the American society for information science*, vol. 45, no. 1, pp. 12–19, 1994.

[115] M. Junker, R. Hoch, and A. Dengel, “On the evaluation of document analysis components by recall, precision, and accuracy,” in *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR’99 (Cat. No. PR00318)*. IEEE, 1999, pp. 713–716.

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

[117] F. Cañas, “Management of agitation in the acute psychotic patient—efficacy without excessive sedation,” *European neuropsychopharmacology*, vol. 17, pp. S108–S114, 2007.

[118] H. Dibeklioğlu, Z. Hammal, Y. Yang, and J. F. Cohn, “Multimodal detection of depression in clinical interviews,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 307–310.

[119] A. Pampouchidou, O. Simantiraki, A. Fazlollahi, M. Pediaditis, D. Manousos, A. Roniotis, G. Giannakakis, F. Meriaudeau, P. Simos, K. Marias *et al.*, “Depression assessment by fusing high and low level features from audio, video, and text,” in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 27–34.

[120] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud *et al.*, “Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition,” in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 3–13.

[121] Z. Du, W. Li, D. Huang, and Y. Wang, “Bipolar disorder recognition via multi-scale discriminative audio temporal representation,” in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 23–30.

[122] N. Kathmann, A. Hochrein, R. Uwer, and B. Bondy, “Deficits in gain of smooth pursuit eye movements in schizophrenia and affective disorder patients and their unaffected relatives,” *American Journal of Psychiatry*, vol. 160, no. 4, pp. 696–702, 2003.

[123] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” 2005.

[124] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[125] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, “Emotion recognition using phog and lpq features,” in *Face and Gesture 2011*. IEEE, 2011, pp. 878–883.

[126] S. Alghowinem, R. Goecke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear, “Cross-cultural detection of depression from nonverbal behaviour,” in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–8.

[127] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, “Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors,” *IEEE Transactions on Affective Computing*, 2016.

[128] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681–685, 2001.

[129] J. W. Davis, “Hierarchical motion history images for recognizing human motion,” in *Proceedings IEEE Workshop on Detection and Recognition of Events in Video*. IEEE, 2001, pp. 39–46.

[130] R. B. Lipton, S. Levin, and P. S. Holzman, “Horizontal and vertical pursuit eye movements, the oculocephalic reflex, and the functional psychoses,” *Psychiatry Research*, vol. 3, no. 2, pp. 193–203, 1980.

[131] L. A. Abel, L. Friedman, J. Jesberger, A. Malki, and H. Meltzer, “Quantitative assessment of smooth pursuit gain and catch-up saccades in schizophrenia and affective disorders,” *Biological psychiatry*, vol. 29, no. 11, pp. 1063–1072, 1991.

[132] I. M. Cameron, J. R. Crawford, A. H. Cardy, S. W. du Toit, K. Lawton, S. Hay, K. Mitchell, S. Sharma, S. Shivaprasad, S. Winning *et al.*, “Psychometric properties of the quick

inventory of depressive symptomatology (qids-sr) in uk primary care," *Journal of psychiatric research*, vol. 47, no. 5, pp. 592–598, 2013.

[133] Y. Li, Y. Xu, M. Xia, T. Zhang, J. Wang, X. Liu, Y. He, and J. Wang, "Eye movement indices in the study of depressive disorder," *Shanghai archives of psychiatry*, vol. 28, no. 6, p. 326, 2016.

[134] J. A. Sweeney, M. H. Strojwas, J. J. Mann, and M. E. Thase, "Prefrontal and cerebellar abnormalities in major depression: evidence from oculomotor studies," *Biological psychiatry*, vol. 43, no. 8, pp. 584–594, 1998.

[135] P. Ekman and W. V. Friesen, "Nonverbal behavior and psychopathology," *The psychology of depression: Contemporary theory and research*, pp. 3–31, 1974.

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

[137] Z. Huang, B. Stasak, T. Dang, K. Wataraka Gamage, P. Le, V. Sethu, and J. Epps, "Staircase regression in oa rvm, data selection and gender dependency in avec 2016," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 19–26.

[138] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, "Detecting depression using vocal, facial and semantic communication cues," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 11–18.

[139] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews." in *LREC*. Citeseer, 2014, pp. 3123–3128.

[140] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[141] S. Scherer, G. Stratou, and L.-P. Morency, "Audiovisual behavior descriptors for depression assessment," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 135–140.

[142] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *Face and gesture 2011*. IEEE, 2011, pp. 298–305.

[143] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–8.

[144] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and ptsd: the effect of gender," *Journal on Multimodal User Interfaces*, vol. 9, no. 1, pp. 17–29, 2015.

[145] S. Vijay, T. Baltrušaitis, L. Pennant, D. Ongür, J. T. Baker, and L.-P. Morency, "Computational study of psychosis symptoms and facial expressions," in *Computing and Mental Health Workshop at CHI*, 2016.

[146] 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 *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 65–72.

[147] A. Królak and P. Strumiłło, "Eye-blink detection system for human–computer interaction," *Universal Access in the Information Society*, vol. 11, no. 4, pp. 409–419, 2012.

[148] A. Rahman, M. Sirshar, and A. Khan, "Real time drowsiness detection using eye blink monitoring," in *Software Engineering Conference (NSEC), 2015 National*. IEEE, 2015, pp. 1–7.

[149] G. F. Wilson, "An analysis of mental workload in pilots during flight using multiple psychophysiological measures," *The International Journal of Aviation Psychology*, vol. 12, no. 1, pp. 3–18, 2002.

[150] J. A. Stern, L. C. Walrath, and R. Goldstein, "The endogenous eyeblink," *Psychophysiology*, vol. 21, no. 1, pp. 22–33. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.1984.tb02312.x

[151] W. O. Lee, E. C. Lee, and K. R. Park, "Blink detection robust to various facial poses," *Journal of neuroscience methods*, vol. 193, no. 2, pp. 356–372, 2010.

[152] F. Song, X. Tan, X. Liu, and S. Chen, "Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients," *Pattern Recognition*, vol. 47, no. 9, pp. 2825–2838, 2014.

[153] B. D. Eddine, F. N. Dos Santos, B. Boulebtateche, and S. Bensaoula, "Eyelsd a robust approach for eye localization and state detection," *Journal of Signal Processing Systems*, vol. 90, no. 1, pp. 99–125, 2018.

[154] L. Zhao, Z. Wang, G. Zhang, Y. Qi, and X. Wang, "Eye state recognition based on deep integrated neural network and transfer learning," *Multimedia Tools and Applications*, vol. 77, no. 15, pp. 19 415–19 438, 2018.

[155] M. Wang, L. Guo, and W.-Y. Chen, "Blink detection using adaboost and contour circle for fatigue recognition," *Computers & Electrical Engineering*, vol. 58, pp. 502–512, 2017.

[156] K. Radlak and B. Smolka, "Blink detection based on the weighted gradient descriptor," in *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013*. Springer, 2013, pp. 691–700.

[157] J. Mohanakrishnan, S. Nakashima, J. Odagiri, and S. Yu, "A novel blink detection system for user monitoring," in *User-Centered Computer Vision (UCCV), 2013 1st IEEE Workshop on*. IEEE, 2013, pp. 37–42.

[158] T. Drutarovsky and A. Fogelton, "Eye blink detection using variance of motion vectors," in *European Conference on Computer Vision*. Springer, 2014, pp. 436–448.

[159] F. M. Sukno, S.-K. Pavani, C. Butakoff, and A. F. Frangi, “Automatic assessment of eye blinking patterns through statistical shape models,” in *International Conference on Computer Vision Systems.* Springer, 2009, pp. 33–42.

[160] L. A. Jeni, J. F. Cohn, and T. Kanade, “Dense 3d face alignment from 2d videos in real-time,” in *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–8.

[161] X. Xiong and F. De la Torre, “Supervised descent method and its applications to face alignment,” in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on.* IEEE, 2013, pp. 532–539.

[162] L. A. Jeni, J. F. Cohn, and T. Kanade, “Dense 3d face alignment from 2d video for real-time use,” *Image and Vision Computing*, vol. 58, pp. 13–24, 2017.

[163] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape, “Offline deformable face tracking in arbitrary videos,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 1–9.

[164] A. Savitzky and M. J. Golay, “Smoothing and differentiation of data by simplified least squares procedures.” *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[165] G. Pan, L. Sun, Z. Wu, and S. Lao, “Eyeblink-based anti-spoofing in face recognition from a generic webcamera,” in *2007 IEEE 11th International Conference on Computer Vision.* IEEE, 2007, pp. 1–8.

[166] A. Fogelton and W. Benesova, “Eye blink detection based on motion vectors analysis,” *Computer Vision and Image Understanding*, vol. 148, pp. 23–33, 2016.

[167] A. R. Bentivoglio, S. B. Bressman, E. Cassetta, D. Carretta, P. Tonali, and A. Albanese, “Analysis of blink rate patterns in normal subjects,” *Movement Disorders*, vol. 12, no. 6, pp. 1028–1034, 1997.

[168] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, “Avec 2014: 3d dimensional affect and depression recognition challenge,” in *Proceedings of the 4th international workshop on audio/visual emotion challenge*, 2014, pp. 3–10.

[169] H. S. Mayberg, A. M. Lozano, V. Voon, H. E. McNeely, D. Seminowicz, C. Hamani, J. M. Schwalb, and S. H. Kennedy, "Deep brain stimulation for treatment-resistant depression," *Neuron*, vol. 45, no. 5, pp. 651–660, 2005.

[170] T. Deckersbach, D. D. Dougherty, and S. L. Rauch, "Functional imaging of mood and anxiety disorders," *Journal of Neuroimaging*, vol. 16, no. 1, pp. 1–10, 2006.

[171] T. A. Brown, P. A. Di Nardo, C. L. Lehman, and L. A. Campbell, "Reliability of dsm-iv anxiety and mood disorders: implications for the classification of emotional disorders." *Journal of abnormal psychology*, vol. 110, no. 1, p. 49, 2001.

[172] W. H. Organization, *The global burden of disease : 2004 update*. Geneva: World Health Organization, 2008.

[173] S. D. Østergaard, S. Jensen, and P. Bech, "The heterogeneity of the depressive syndrome: when numbers get serious," *Acta Psychiatrica Scandinavica*, vol. 124, no. 6, pp. 495–496, 2011.

[174] M. P. Caligiuri and J. Ellwanger, "Motor and cognitive aspects of motor retardation in depression," *Journal of affective disorders*, vol. 57, no. 1-3, pp. 83–93, 2000.

[175] M. Hamilton, "A rating scale for depression," *Journal of neurology, neurosurgery, and psychiatry*, vol. 23, no. 1, p. 56, 1960.

[176] W. HO, "Depression and other common mental disorders," *Global Health Estimates*, 2017.

[177] G. Hasler, W. C. Drevets, H. K. Manji, and D. S. Charney, "Discovering endophenotypes for major depression," *Neuropsychopharmacology*, vol. 29, no. 10, pp. 1765–1781, 2004.

[178] M. Åsberg, "Neurotransmitters and suicidal behavior: The evidence from cerebrospinal fluid studies a," *Annals of the New York Academy of Sciences*, vol. 836, no. 1, pp. 158–181, 1997.

[179] J. J. Mann, "Neurobiology of suicidal behaviour," *Nature Reviews Neuroscience*, vol. 4, no. 10, pp. 819–828, 2003.

[180] G. MacQueen and T. Frodl, “The hippocampus in major depression: evidence for the convergence of the bench and bedside in psychiatric research?” *Molecular psychiatry*, vol. 16, no. 3, pp. 252–264, 2011.

[181] T. Frodl, C. Schüle, G. Schmitt, C. Born, T. Baghai, P. Zill, R. Bottlender, R. Rupprecht, B. Bondy, M. Reiser *et al.*, “Association of the brain-derived neurotrophic factor val66met polymorphism with reduced hippocampal volumes in major depression,” *Archives of General Psychiatry*, vol. 64, no. 4, pp. 410–416, 2007.

[182] A. Steiger and M. Kimura, “Wake and sleep eeg provide biomarkers in depression,” *Journal of psychiatric research*, vol. 44, no. 4, pp. 242–252, 2010.

[183] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, “Multimodal assistive technologies for depression diagnosis and monitoring,” *Journal on Multimodal User Interfaces*, vol. 7, no. 3, pp. 217–228, 2013.

[184] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, “Detecting depression from facial actions and vocal prosody,” in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–7.

[185] J. F. Cohn and F. De la Torre, “Automated face analysis for affective computing.” 2015.

[186] M. Hamilton, “Arating scalefordepression. journalofneurology,” *Neurosurgery, and Psychiatry*, vol. 23, p. 5642, 1960.

[187] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh, “An inventory for measuring depression,” *Archives of general psychiatry*, vol. 4, no. 6, pp. 561–571, 1961.

[188] M. Hamilton, “Assessment of change in psychiatric state by means of rating scales,” *Proceedings of the Royal Society of Medicine*, vol. 59, no. Suppl 1, p. 10, 1966.

[189] ——, “Development of a rating scale for primary depressive illness,” *British journal of social and clinical psychology*, vol. 6, no. 4, pp. 278–296, 1967.

[190] ——, “Standardised assessment and recording of depressive symptoms.” *Psychiatria, Neurologia, Neurochirurgia*, vol. 72, no. 2, pp. 201–205, 1969.

[191] ——, “Rating depressive patients.” *The Journal of clinical psychiatry*, 1980.

[192] R. Nuevo, V. Lehtinen, P. M. Reyna-Liberato, and J. L. Ayuso-Mateos, “Usefulness of the beck depression inventory as a screening method for depression among the general population of finland,” *Scandinavian journal of public health*, vol. 37, no. 1, pp. 28–34, 2009.

[193] 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.* Springer, 2009, pp. 7–35.

[194] H. Ellgring, *Non-verbal communication in depression.* Cambridge University Press, 2007.

[195] G. McIntyre, R. Göcke, M. Hyett, M. Green, and M. Breakspear, “An approach for automatically measuring facial activity in depressed subjects,” in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops.* IEEE, 2009, pp. 1–8.

[196] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, “Avec 2016: Depression, mood, and emotion recognition workshop and challenge,” in *Proceedings of the 6th international workshop on audio/visual emotion challenge.* ACM, 2016, pp. 3–10.

[197] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, “Avec 2013: the continuous audio/visual emotion and depression recognition challenge,” in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge.* ACM, 2013, pp. 3–10.

[198] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, “Multimodal prediction of affective dimensions and depression in

human-computer interactions," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 33–40.

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

[200] H. Pérez Espinosa, H. J. Escalante, L. Villaseñor-Pineda, M. Montes-y Gómez, D. Pinto-Avedaño, and V. Reyez-Meza, "Fusing affective dimensions and audio-visual features from segmented video for depression recognition: Inaoe-buap's participation at avec'14 challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 49–55.

[201] D. Kupfer and F. G. Foster, "Interval between onset of sleep and rapid-eye-movement sleep as an indicator of depression," *The Lancet*, vol. 300, no. 7779, pp. 684–686, 1972.

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

[203] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*. Springer, 1995, pp. 23–37.

[204] R. Meir and G. Rätsch, "An introduction to boosting and leveraging," in *Advanced lectures on machine learning*. Springer, 2003, pp. 118–183.

[205] N. Cristianini, J. Shawe-Taylor *et al.*, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[206] V. Vapnik, "Statistical learning theory wiley-interscience," *New York*, 1998.

[207] R. E. Schapire and Y. Freund, "Boosting: Foundations and algorithms," *Kybernetes*, 2013.

[208] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, “Automated depression diagnosis based on deep networks to encode facial appearance and dynamics,” *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 578–584, 2017.

[209] H. Kaya, F. Çilli, and A. A. Salah, “Ensemble cca for continuous emotion prediction,” in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 19–26.

[210] D. Ebert, R. Albert, G. Hammon, B. Strasser, A. May, and A. Merz, “Eye-blink rates and depression: Is the antidepressant effect of sleep deprivation mediated by the dopamine system?” *Neuropsychopharmacology*, vol. 15, no. 4, pp. 332–339, 1996.

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

[212] J. Kittler, M. Hatef, R. Duin, and J. Matas, “On combining classifiers, ieee transactions on pattern analysis and mach,” 1998.

[213] B. Klimova and K. Kuca, “Speech and language impairments in dementia,” *Journal of Applied Biomedicine*, vol. 14, no. 2, pp. 97–103, 2016.

[214] Z. S. Khachaturian, “Alzheimer’s & dementia: the journal of the alzheimer’s association,” *Alzheimer’s & dementia: the journal of the Alzheimer’s Association*, vol. 4, no. 5, p. 315, 2008.

[215] A. Alzheimer’s, “Alzheimer’s disease facts and figures, alzheimer’s dement,” *J. Alzheimer’s Assoc*, vol. 11, no. 3, 2015.

[216] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack Jr, J. Kaye, T. J. Montine *et al.*, “Toward defining the preclinical stages of alzheimer’s disease: Recommendations from the national institute on aging-alzheimer’s association workgroups on diagnostic guidelines for alzheimer’s disease,” *Alzheimer’s & dementia*, vol. 7, no. 3, pp. 280–292, 2011.

[217] C. Iadecola, “The pathobiology of vascular dementia,” *Neuron*, vol. 80, no. 4, pp. 844–866, 2013.

[218] D. J. Blackburn, S. Wakefield, M. F. Shanks, K. Harkness, M. Reuber, and A. Venneri, "Memory difficulties are not always a sign of incipient dementia: a review of the possible causes of loss of memory efficiency," *British medical bulletin*, vol. 112, no. 1, pp. 71–81, 2014.

[219] A. Alberdi, A. Aztiria, and A. Basarab, "On the early diagnosis of alzheimer's disease from multimodal signals: A survey," *Artificial intelligence in medicine*, vol. 71, pp. 1–29, 2016.

[220] D. Jolley, S. Benbow, and M. Grizzell, "Memory clinics," *Postgraduate Medical Journal*, vol. 82, no. 965, pp. 199–206, 2006.

[221] C. Laske, H. R. Sohrabi, S. M. Frost, K. López-de Ipiña, P. Garrard, M. Buscema, J. Dauwels, S. R. Soekadar, S. Mueller, C. Linnemann *et al.*, "Innovative diagnostic tools for early detection of alzheimer's disease," *Alzheimer's & Dementia*, vol. 11, no. 5, pp. 561–578, 2015.

[222] R. Camicioli, "Diagnosis and differential diagnosis of dementia," *Dementia*, pp. 1–13, 2014.

[223] E. E. Camargo, "Brain spect in neurology and psychiatry," *Journal of Nuclear Medicine*, vol. 42, no. 4, pp. 611–623, 2001.

[224] C. Elsey, P. Drew, D. Jones, D. Blackburn, S. Wakefield, K. Harkness, A. Venneri, and M. Reuber, "Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics," *Patient Education and Counseling*, vol. 98, no. 9, pp. 1071–1077, 2015.

[225] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "“mini-mental state": a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[226] J. C. Anthony, L. LeResche, U. Niaz, M. R. Von Korff, and M. F. Folstein, "Limits of the 'mini-mental state'as a screening test for dementia and delirium among hospital patients," *Psychological medicine*, vol. 12, no. 2, pp. 397–408, 1982.

[227] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[228] S. Hodge and E. Hailey, "English national memory clinics audit report," *London: Royal College of Psychiatrists*, vol. 20132013, 2013.

[229] S. Bell, K. Harkness, J. Dickson, and D. Blackburn, "A diagnosis for£ 55: what is the cost of government initiatives in dementia case finding," *Age and ageing*, vol. 44, no. 2, pp. 344–345, 2015.

[230] A. Ladas, C. Frantzidis, P. Bamidis, and A. B. Vivas, "Eye blink rate as a biological marker of mild cognitive impairment," *International Journal of Psychophysiology*, vol. 93, no. 1, pp. 12–16, 2014.

[231] B. Mirheidari, D. Blackburn, M. Reuber, T. Walker, and H. Christensen, "Diagnosing people with dementia using automatic conversation analysis," in *Proceedings of interspeech*. ISCA, 2016, pp. 1220–1224.

[232] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "An avatar-based system for identifying individuals likely to develop dementia," in *Interspeech 2017*. ISCA, 2017, pp. 3147–3151.

[233] ——, "Toward the automation of diagnostic conversation analysis in patients with memory complaints," *Journal of Alzheimer's Disease*, vol. 58, no. 2, pp. 373–387, 2017.

[234] D. Jones, P. Drew, C. Elsey, D. Blackburn, S. Wakefield, K. Harkness, and M. Reuber, "Conversational assessment in memory clinic encounters: interactional profiling for differentiating dementia from functional memory disorders," *Aging & Mental Health*, vol. 20, no. 5, pp. 500–509, 2016.

[235] A. J. Larner, "Addenbrooke's cognitive examination-revised (ace-r) in day-to-day clinical practice," *Age and ageing*, vol. 36, no. 6, pp. 685–686, 2007.

[236] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: the gad-7," *Archives of internal medicine*, vol. 166, no. 10, pp. 1092–1097, 2006.

[237] J. C. Raven, "Guide to using the coloured progressive matrices." 1958.

[238] J. Ridley, "Stroop. studies of interference in serial verbal reactions," *Journal of Experimental Psychology: Generak*, vol. 121, no. 1, p. 15, 1992.

[239] E. De Renzi and P. Faglioni, "Normative data and screening power of a shortened version of the token test," *Cortex*, vol. 14, no. 1, pp. 41–49, 1978.

[240] D. Wechsler, "Wechsler adult intelligence scale–," 1955.

[241] V. Kazemi and S. Josephine, "One millisecond face alignment with an ensemble of regression trees," in *27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, United States, 23 June 2014 through 28 June 2014.* IEEE Computer Society, 2014, pp. 1867–1874.

[242] F. Timm and E. Barth, "Accurate eye centre localisation by means of gradients." *Visapp*, vol. 11, pp. 125–130, 2011.

[243] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2079–2107, 2010.

[244] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[245] T. D. Wilcockson, D. Mardanbegi, B. Xia, S. Taylor, P. Sawyer, H. W. Gellersen, I. Leroi, R. Killick, and T. J. Crawford, "Abnormalities of saccadic eye movements in dementia due to alzheimer's disease and mild cognitive impairment," *Aging (Albany NY)*, vol. 11, no. 15, p. 5389, 2019.

[246] R. K. Brown, N. I. Bohnen, K. K. Wong, S. Minoshima, and K. A. Frey, “Brain pet in suspected dementia: patterns of altered fdg metabolism,” *Radiographics*, vol. 34, no. 3, pp. 684–701, 2014.

[247] A. Villarejo and V. Puertas-Martín, “Usefulness of short tests in dementia screening,” *Neurología (English Edition)*, vol. 26, no. 7, pp. 425–433, 2011.

[248] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, “Avec 2011–the first international audio/visual emotion challenge,” in *International Conference on Affective Computing and Intelligent Interaction.* Springer, 2011, pp. 415–424.

[249] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, “Avec 2012: the continuous audio/visual emotion challenge,” in *Proceedings of the 14th ACM international conference on Multimodal interaction.* ACM, 2012, pp. 449–456.

[250] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, “Avec 2015: The 5th international audio/visual emotion challenge and workshop,” in *Proceedings of the 23rd ACM international conference on Multimedia.* ACM, 2015, pp. 1335–1336.

[251] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, “Avec 2017: Real-life depression, and affect recognition workshop and challenge,” in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge.* ACM, 2017, pp. 3–9.

[252] M. Leboyer and D. J. Kupfer, “Bipolar disorder: new perspectives in health care and prevention,” *The Journal of clinical psychiatry*, vol. 71, no. 12, p. 1689, 2010.

[253] M. Sajatovic, “Bipolar disorder: disease burden,” *Am J Manag Care*, vol. 11, no. 3 Suppl, pp. S80–84, 2005.

[254] C. J. Murray, T. Vos, R. Lozano, M. Naghavi, A. D. Flaxman, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla *et al.*, “Disability-adjusted life years (dalys) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010,” *The lancet*, vol. 380, no. 9859, pp. 2197–2223, 2012.

[255] I. E. Bauer, J. C. Soares, S. Selek, and T. D. Meyer, “The link between refractoriness and neuroprogression in treatment-resistant bipolar disorder,” in *Neuroprogression in Psychiatric Disorders*. Karger Publishers, 2017, vol. 31, pp. 10–26.

[256] E. Çiftçi, H. Kaya, H. Güleç, and A. A. Salah, “The turkish audio-visual bipolar disorder corpus,” in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. IEEE, 2018, pp. 1–6.

[257] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, “The interspeech 2010 paralinguistic challenge,” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[258] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: theory and applications,” *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[259] R. Rosipal and N. Krämer, “Overview and recent advances in partial least squares,” in *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*. Springer, 2005, pp. 34–51.

[260] L. Yang, Y. Li, H. Chen, D. Jiang, M. C. Oveneke, and H. Sahli, “Bipolar disorder recognition with histogram features of arousal and body gestures,” in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 15–21.

[261] Z. S. Syed, K. Sidorov, and D. Marshall, “Automated screening for bipolar disorder from audio/visual modalities,” in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 39–45.

[262] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in opensmile, the munich open-source multimedia feature extractor,” in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[263] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[264] C. Berchio, C. Piguet, C. M. Michel, P. Cordera, T. A. Rihs, A. G. Dayer, and J.-M. Aubry, “Dysfunctional gaze processing in bipolar disorder,” *Neuroimage: clinical*, vol. 16, pp. 545–556, 2017.

[265] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[266] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, “API design for machine learning software: experiences from the scikit-learn project,” in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[267] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[268] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, “Deep face recognition.” in *bmvc*, vol. 1, no. 3, 2015, p. 6.

[269] S. Zhang, C. Zhang, and Q. Yang, “Data preparation for data mining,” *Applied artificial intelligence*, vol. 17, no. 5-6, pp. 375–381, 2003.

[270] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database forstudying face recognition in unconstrained environments,” in *Workshop on faces in’Real-Life’Images: detection, alignment, and recognition*, 2008.

[271] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[272] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[273] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, “How fast are the leaked facial expressions: The duration of micro-expressions,” *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013.

[274] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv preprint arXiv:1502.03167*, 2015.

[275] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[276] L. Chao, J. Tao, M. Yang, and Y. Li, “Multi task sequence learning for depression scale prediction from video,” in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 526–531.

[277] T. R. Almaev and M. F. Valstar, “Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition,” in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 356–361.

[278] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[279] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, “Yaafe, an easy to use and efficient audio feature extraction software.” in *ISMIR*, 2010, pp. 441–446.

[280] 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 *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 41–48.

[281] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch,” *arXiv preprint arXiv:1411.7923*, 2014.

[282] M. Sidorov and W. Minker, “Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach,” in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 81–86.

[283] G. W. Ross and J. D. Bowen, “The diagnosis and differential diagnosis of dementia.” *The Medical Clinics of North America*, vol. 86, no. 3, pp. 455–476, 2002.

[284] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, “An image-based deep spectrum feature representation for the recognition of emotional speech,” in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 478–484.

[285] X. Xing, B. Cai, Y. Zhao, S. Li, Z. He, and W. Fan, “Multi-modality hierarchical recall based on gbdts for bipolar disorder classification,” in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 31–37.

[286] Y. Gong and C. Poellabauer, “Topic modeling based multi-modal depression detection,” in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 69–76.

[287] J. Z. Huang, “An introduction to statistical learning: With applications in r by gareth james, trevor hastie, robert tibshirani, daniela witten,” 2014.

[288] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system in proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (acm, san francisco, california, usa, 2016), 785–794. isbn: 978-1-4503-4232-2. doi: 10.1145/2939672.2939785,” 2016.

[289] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” *arXiv preprint arXiv:1611.08050*, 2016.

[290] P. Pudil, J. Novovičová, and J. Kittler, “Floating search methods in feature selection,” *Pattern recognition letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[291] N. Jaques, D. McDuff, Y. L. Kim, and R. Picard, “Understanding and predicting bonding in conversations using thin slices of facial expressions and body language,” in *International Conference on Intelligent Virtual Agents.* Springer, 2016, pp. 64–74.

[292] M. Kane and G. Terry, “Dementia 2015: aiming higher to transform lives,” *London: Alzheimer’s Society*, 2015.

[293] K. Hawton, C. C. i Comabella, C. Haw, and K. Saunders, “Risk factors for suicide in individuals with depression: a systematic review,” *Journal of affective disorders*, vol. 147, no. 1-3, pp. 17–28, 2013.

[294] S. Al-gawwam and M. Benaissa, “Robust eye blink detection based on eye landmarks and savitzky–golay filtering,” *Information*, vol. 9, no. 4, p. 93, 2018.

[295] S. Al-Gawwam and M. Benaissa, “Eye blink detection using facial features tracker,” in *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*. ACM, 2017, pp. 27–30.

[296] S. Al-gawwam and M. Benaissa, “Depression detection from eye blink features,” in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2018, pp. 388–392.

[297] T. Althoff, K. Clark, and J. Leskovec, “Large-scale analysis of counseling conversations: An application of natural language processing to mental health,” *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 463–476, 2016.

[298] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, “Natural language processing in mental health applications using non-clinical texts,” *Natural Language Engineering*, vol. 23, no. 5, pp. 649–685, 2017.