The Effectiveness of Facial Cues for Automatic Detection of Cognitive Impairment

Using In-the-wild Data

Fatimah Alzahrani

Supervisors: Prof. Heidi Christensen and Dr. Steve Maddock

Department of Computer Science
The University of Sheffield

This dissertation is submitted for the degree of
Doctor of Philosophy

Speech and Hearing Research Group November 2023

I would like to dedicate this thesis to
My beloved parents, Ahmed and Badryah, the source of my strength, inspiration and happiness and my reasons to keep going
My sisters and brothers - for their love and support

Declaration

I hereby declare that this dissertation is of my own work, except where I specifically referred to the works done by the other authors in the text. The contents of the study are original and have not been submitted for any other awards, qualifications, or degrees in universities. Parts of the findings of this study have already been published in a conference proceedings.

Fatimah Alzahrani
November 2023

Acknowledgements

This PhD has been a life-changing experience for me by helping me to face many challenges and achieve my aims, which would not have been possible without the invaluable support from different people (in my personal and professional lives).

First and foremost, I would like to express my deepest gratitude to Allah, without whom nothing is possible. Then, I would like to thank my supervisors, Prof. Heidi Christensen and Dr. Steve Maddock, who, without them, this thesis would not have been accomplished. Thank you for everything you have done for me from the beginning till the end, including your guidance, patience and endless support and encouragement of me and my ideas. No words can express my appreciation for your outstanding mentorship, kindness, and advice, which made me a much better student and researcher. Working with this exceptional supervisory team has been an honour and privilege. I would also like to extend my gratitude to my panel members, Dr. Yoshi Gotoh and Dr. Vita Lanfranchi, for their helpful comments and suggestions regarding my work through each panel meeting.

My sincere thanks go to my colleagues in the Speech and Hearing Lab, who have always been a continuous source of knowledge and friendship, especially Rabab, Dalia, Areej, Ablah, Jack, Gerardo, Zhengjun, Zehai, Nathan, Hend, Bahman, Yilin, Jisi and Wanli.

I would not have gotten through this journey without my family. Thank you for keeping me strong through this journey and supporting me, especially my parents, the light of my eyes, who have been there for me every step of the way during my study abroad in both degrees MSc and PhD, especially my father Ahmed who always stands by my side and answers 'Yes' for everything I want until I reach the person who I am now. I know you have given my siblings and me everything you can through all these years, and my only hope is that I can be the daughter you wish.

وَاذْكُرِ الأَبَ العَطُوفَ المُرْشِدَا   إنَه لَيثٌ بدَارٍ كالعَرِينْ
بَاكِراً فِي سَعْيِهِ مِثْلَ الصُقورْ   جَاعِلاً مِنْ بَيتهِ الحِصْنَ الحَصِينْ
مُتْعَبَاً فِي هَمِه لا يَشْتَكِي   فِي نَشَاطٍ دَائِمٍ لا يَسْتَكِينْ

My dear mother Badryah, I can never thank you enough. I feel I have been blessed by having the most kind and loving-hearted mother in the world.

أمّي وإن طالَ الحديث بها   فلا شِعرٌ يُوَفّيها ولا الأَقلامُ

My dearest father and mother, I am so grateful for being your daughter and feel privileged to have been raised by you. Thank you for your endless love, support and prayers that underpinned my persistence on this journey and made the completion of this thesis possible. I would have never wished to have any parents instead of you.

I also feel incredibly blessed for the love and support I continuously receive from my sisters Noura, Arwa, Rasha, and Mashael; and brothers Saeed, Mohammed, Ibrahim, and Yousef. I owe a special thanks to my nieces and nephews Jomana, Seba, Dana, Faisal, Ahmed and Rayan. Thank you all for all your loving actions, warm welcomes whenever I am home, and incredible sense of humour.

I am hugely grateful to my friends, my second family in this journey, for always being there for me and for all their shown encouragement and love, especially Nadia, Asma, Ohoud, Areej and Amal. Thank you from the bottom of my heart for your support and love and for keeping me strong throughout this journey. I have been blessed with a very loving and supportive family and friends, who are a priceless gift.

Finally, I thankfully acknowledge the financial funding for this work from Umm Al-Qura University and the Saudi Ministry of Education, to which I am grateful.

Abstract

The development of automatic methods for the early detection of cognitive impairment (CI) has attracted much research interest due to its crucial role in helping people get suitable treatment or care. People with CI may experience various changes in their facial cues, such as eye blink rate and head movement.

This thesis aims to investigate the use of facial cues to develop an automatic system for detecting CI using in-the-wild data. Firstly, the 'in-the-wild data' term is defined, and associated challenges are identified by analysing datasets used in previous work. In-the-wild data can affect the reliability of the performance of state-of-the-art approaches. Second, this thesis investigates the automatic detection of neurodegenerative disorder, mild cognitive impairment and functional memory disorder, showing the applicability of detecting health conditions with similar symptoms.

Then, a novel multiple thresholds (MTs) approach for detecting an eye blink rate feature is introduced. This approach addresses in-the-wild data challenges by generating multiple thresholds, resulting in a vector of blink rates for each participant. Then, the feasibility of this feature in detecting CI is examined. Other features considered are head turn rate, head turn statistical features, head movement statistical features and low-level features. The results show that these facial features significantly distinguish different health conditions.

Next, the MTs approach is validated on a public dataset for people with depression, achieving results comparable to related work. An evaluation of the system developed is then carried out on a larger and more varied dataset for people with CI, which results in extending the MTs approach to be participant-dependent. The findings show that it is feasible to develop an automatic system for detecting CI by analysing facial cues using an in-the-wild dataset. In conclusion, this thesis makes promising progress in improving the detection of CI through low-cost and non-invasive approaches alternative to current expensive assessments.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

300VW	300 videos in the wild
300W	300 Faces in the Wild
AA	attended alone
AAM	active appearance model
ABD	autonomus blink dataset
AD	Alzheimer's disease
AFLW	Annotated Facial Landmarks in the Wild
AFW	Aannotated Faces in the Wild
AVEC 2014	audio-visual emotion challenge 2014
AW	attended with
BIC	bayesian information criterion
CI	cognitive impairment
CNNs	convolutional neural networks
CT	computed tomography
CV	cross-validation
DA	dopamine activity
D	dementia

DDF	distracted driver face
DG	depressed group
DLB	dementia with lewy bodies
DT	decision trees
EAR	eye aspect ratio
EBR	eye blink rate
EM	expectation maximization
FAUs	facial action units
FDG-PET	fluorodeoxyglucose positron emission tomography
FMD	functional memory disorders
FN	false negative
FP	false positive
fps	frames per second
FTD	Frontotemporal dementia
GMM	Gaussian mixture model
HC	healthy controls
HTSF	head movement statistical features
HTR	head turn rate
HTSF	head turn statistical features
IJB-A	IARPA Janus benchmark
IOU	intersection over union
IQR	interquartile range
KNN	k-nearest neighbours
LB	lower boundary

LFPW	Labelled Face Parts in the Wild
LFW	Labelled Faces in the Wild
LLFs	low-level features
LR	logistic regression
MAE	mean absolute error
MCI	mild cognitive impairment
MLE	maximum likelihood estimation
MMSE	minimal mental status examination
MoCA	Montreal cognitive assessment
MP	memory problems
MRI	magnetic resonance imaging
MTs	multiple thresholds
NDG	non-depressed group
ND	neurodegenerative disorder
NDS	naturalistic driving study
NND	number needed to diagnose
NNM	number needed to misdiagnose
NNP	number needed to predict
NPH	normal pressure hydrocephalus
OAP	over all participants
PD	Parkinson’s disease
PD	participant-dependent
PSP	progressive supranuclear palsy
RFECV	recursive feature elimination cross-validation

RMSE	root mean squared error
RU-FACS	rochester/UCSD-facial action coding system
SD	standard deviation
SM	state machine
SPECT	single-photon emission computed tomography
SVM	support vector machine
SVR	support vector regression
SW	Shapiro-Wilk
TN	true negative
TP	true positive
UB	upper boundary
VaD	vascular dementia
YEC	Youtube eye-state classification
YMRS	young mania rating scale

Chapter 1

Introduction

"Mental health affects every aspect of your life. It's not just this neat little issue you can put into a box."
— Shannon Purser

1.1 Motivation

As people age, they can exhibit cognitive changes that are common and related to normal ageing, such as difficulty recalling the names of people or places, taking longer to process new information, and occasionally forgetting appointments or events (Ahmad et al., 2013; Smith et al., 2022). When these changes cause a serious decline in certain cognitive functions beyond what is expected from a person's age, it can be diagnosed as cognitive decline. Cognitive decline can be categorised into three different levels of cognitive decline: cognitive impairment (CI), mild cognitive impairment (MCI) and dementia (van de Mortel et al., 2021).

CI is a broad term referring to any decline in cognitive function, which can range from mild to severe and may progress to more advanced stages (Allan et al., 2017). Several changes in mood, emotion, or behaviour can occasionally precede the progression of CI. MCI, which is a specific type of CI, is the transitional stage between normal ageing and dementia. People with MCI can experience a severe decline in memory, thinking and remembering important events that are noticeable to themselves and their families, but it does not affect their activities of daily living (Cermakova et al., 2020; Horackova et al., 2019; UCSF Memory and Aging Center, 2020). Although those people are at high risk of developing dementia, others may remain stable or improve (Petersen et al., 2018). Dementia is a progression of MCI and aggravates over time, causing a significant decline in cognition that affects several cognitive functions, such as memory, attention, learning capacity and language and causes a lack of independence in activities of daily living (Kirk and Berntsen, 2018; UCSF Memory and

Aging Center, 2020). Dementia can be caused by different conditions, such as Alzheimer's disease, which is the most common one, Vascular dementia and Parkinson's disease. People in the later stage of dementia might experience severe memory loss, communication issues, significant changes in verbal and non-verbal behaviours, and even their ability to carry out their activities of daily living. Consequently, people with severe dementia could be reliant on others 24 hours a day for their care. In this thesis, both MCI and dementia will be referred to as 'CI'.

Worldwide, CI is regarded as the seventh most frequent cause of death and one of the major causes of disability for older people (WHO, 2022). The impact of CI can be physical, psychological, social and economic. This effect is not only restricted to the people who live with CI but also affects their families, carers and society. WHO (2022) reported that there are 55 million people living with CI and that more than 60% of them live in developing countries. This number is predicted to increase significantly to 78 million in 2030 and 139 million in 2050.

Diagnosing CI in its early stage is essential. There is no reliable test to diagnose CI due to the overlap between the symptoms and the ageing factor (Travis et al., 1985). In order to help make the diagnosis, doctors take a clinical history and a collateral history and examine the patients. They undertake several other tests, most commonly structural brain scans (computed tomography (CT) or magnetic resonance imaging (MRI)). If the diagnosis is not clear, extra tests will be conducted, such as functional scans (fluorodeoxyglucose positron emission tomography (FDG-PET) and single-photon emission computed tomography (SPECT)) and cerebrospinal fluid analysis of protein biomarkers of Alzheimer's disease (tau and amyloid) (Alberdi et al., 2016). These tests and procedures require multiple appointments for the patient, are costly and take a lot of time and effort.

There are also cognitive tools that can be used by a neurologist based on the patient's condition to assess the CI severity, the most common of which are minimal mental status examination (MMSE) (Folstein et al., 1975) and the Montreal cognitive assessment (MoCA) (Nasreddine et al., 2005). MMSE is used to evaluate orientation to time and space, attention, naming objects, calculation, repeating and remembering words. MoCA is a screening test originally developed to identify MCI. It assesses short-term memory, visuospatial abilities (e.g., a clock-drawing task), attention, concentration, working memory (e.g., calculation), orientation and language (repeating words and naming animals).

In summary, there is no single clinical test or assessment that can be used to diagnose CI. Some patients may have similar symptoms but do not have a neurological disease, which makes a proper diagnosis quite complicated (Gifford and Cummings, 1999). Traditional tests, including but not limited to MMSE, MoCA and interviewing the patient for testing the short-term and long-term memory, are dependent on the ability of the patients to remember and interpret correctly their experiences. In addition, older testing methods are highly dependent on the patients' willingness to share their real emotions, symptoms, and cognition.

To remedy these shortcomings, clinicians have carried out research that employs the use of facial behaviour to diagnose CI (Bouchard and Rossor, 2007). Here, a conversation between the clinician and the patient is recorded using audio and video and then analysed to assess different cues. They found that people with CI may experience a variety of facial behaviour that can help to determine their cognitive state. Some facial cues may be seen in people with CI, including reduced emotional expressiveness (smiling less frequently) and repetitive facial behaviour (eye blinking excessively and head movement). Some clinicians found that people with CI tend to have a higher blink rate than healthy controls (HC) (Jongkees and Colzato, 2016; Ladas et al., 2014) and tend to turn their heads more when asked a question (Fukui et al., 2011; Larner, 2018; Soysal et al., 2017). However, studying such behaviours in a clinic is relatively complex, costly in terms of time and effort and requires professional neurological expertise. In addition, clinicians can face difficulties in keeping track of these cues from recorded videos and analysing them manually for every participant taking into account the high number of people who come to the clinic, which can take several weeks or months. Therefore, a cost-effective tool is needed for the detection of CI.

Many studies have attempted to investigate the automatic detection of CI using acoustic and speech features (Mirheidari et al., 2019; Petti et al., 2020; Walker et al., 2023). However, few studies have investigated the automatic detection of CI using facial cues and achieved good results. When such systems are moved from research into real-world deployment, the performance of such systems may not be reliable because these studies used data recorded in a lab-controlled environment. In order to develop such a tool that overcomes the above-mentioned issues, it is very important to use data recorded in the wild– Chapter 3 will discuss in-the-wild data in more detail. Such data includes many challenges that affect the performance of the available state-of-the-art approaches. Developing such a tool using in-the-wild data can help to open doors to developing a home-based application in the future. In addition, it may help to encourage many patients suspected of having memory problems to do the session at home, which is a comfortable environment for them. It can also help to reduce the pressure on memory clinics. Moreover, it may help clinicians to save time and effort in data collection, monitoring patient health, and diagnosing diseases.

Specific facial cues, eye blink rate (EBR) and head movement, convey important information in assessing cognitive state. Nevertheless, investigating them using computer vision approaches remains under-explored for CI detection. This thesis proposes an automatic system for detecting CI by analysing facial features, including handling data recorded in the wild to reflect the approach's performance when the assessment is made at home via an application.

1.2 Thesis Aim

Increased attention has been paid to assistive technology and healthcare using computer vision and machine learning techniques. However, few related works have explored the integration of both fields to construct a clinical diagnostic application for people with memory problems, such as CI (Gao et al., 2018; Leo et al., 2020; Zolfaghari et al., 2022). This thesis is designed to demonstrate the benefits of this integration among different research fields by using a robust computer vision toolkit to extract the facial landmarks of the participants, calculating several visual features and then analysing them using machine learning classifiers to give a final decision on a participant's diagnostic group. This work can be considered an important step towards helping doctors to use advanced technologies to save time and effort.

Diagnosing CI via video recordings is considered challenging due to the difficulty of having access to suitable data for the extraction of visual features from the patient and applying suitable machine learning techniques to detect the participant's group. Two studies have concentrated on finding ways to automatically detect CI from HC when the CI group includes people with MCI and Alzheimer's disease (Tanaka et al., 2017, 2016). This detection often involves using different modalities, such as language, speech, one facial expression (smile), and machine learning classifiers. Later, Tanaka et al. (2019) explored the performance of using facial action units (FAUs) for detecting CI. These studies achieved good results.

Although some studies have attempted to investigate the visual modality, they used data recorded in a lab-controlled environment. Their studies have been built on data that does not reflect home-recording scenarios, where people at home feel more comfortable as they are not provided with strict instructions about where to sit, how to sit, where to look, and what device to use. These challenges can pose issues for applications developed based on data recorded in a lab-controlled environment in terms of their performance and reliability.

This research aims to investigate the use of facial cues to develop an automatic system for detecting CI using in-the-wild data. In order to achieve the aim of this research, three research questions will be addressed, which are outlined as follows:

RQ.1 There is increasing interest in using in-the-wild datasets that contain many challenges regarding the environment of the recordings and the participants' behaviour to train and evaluate their approaches (Belhumeur et al., 2013; Huang et al., 2008; Shen et al., 2015; Zafeiriou et al., 2017). Therefore, several datasets have been collected as in-the-wild data. However, there is no previous work examined what kind of challenges should be included in this type of data. In addition, there is no in-the-wild dataset for people with health conditions. It is, therefore, important first to answer this question: **What are the kinds of challenges and diversity that should be included in those datasets to make them as representative as possible of real-world environments?**

RQ.2 A few studies have investigated the use of the visual modality in the automatic detection of CI using data recorded in a lab-controlled environment. However, research to date has not yet used in-the-wild data for healthcare research. Therefore, the next research question is **How can facial features be automatically detected in a robust way for in-the-wild data?**

RQ.3 According to the medical literature, patients with CI show a higher blink rate than HC. Clinicians have also found that patients suspected of having CI are more likely to turn their heads to the left or right as an indication of confusion and problems in memory. Some studies have investigated the use of facial movements, such as facial expressions and FAUs, in the automatic detection of CI. However, no work has automatically investigated the use of the EBR and head movement features for CI detection. Thus, **how useful are eye blink rate and head movement for CI detection?**

To answer these questions, this research will be carried out using two datasets recorded in the wild for people with CI: the IVA_{18} and the IVA_{52} datasets. The IVA_{18} dataset consists of 18 participants split equally into 6 neurodegenerative disorder (ND), 6 MCI, and 6 functional memory disorders (FMD). The IVA_{52} dataset includes 52 participants, divided into 11 ND, 10 MCI, 8 FMD and 23 HC. These datasets are described in more detail in Chapter 3.

1.3 Thesis Contributions

The contributions of this thesis can be summarised as follows:

Contribution 1: Analysing the challenges associated with 'in-the-wild data' through an examination of the challenges observed in existing datasets claiming to exhibit in-the-wild properties.

Several attempts have been made to collect in-the-wild data due to its importance in developing state-of-the-art approaches with reliable performance for face detection and facial landmark tracking (Belhumeur et al., 2013; Huang et al., 2008; Shen et al., 2015; Zafeiriou et al., 2017). In-the-wild data can expose the true challenges in real-world scenarios, such as background noise and having people with variations in gender, culture, colour, clothes and eyeglasses. Although extensive research has been carried out on collecting in-the-wild data, no single study exists which provides the kind of challenges that should be included in such a dataset. This research is the first to provide a summary of what the term ‘in-the-wild data’ can cover, explaining the challenges involved regarding the recording environment, participants’ demographics, look and behaviour, and devices used. It also highlights the barriers and the importance of collecting such data, taking into account the common challenges that would be included in data to be considered in the wild, as described in Chapter 3. This contribution addresses research question RQ.1.

Contribution 2: Investigating the use of in-the-wild data for people with CI.

In-the-wild video recordings for people with CI would enable the development of an approach that handles such data to achieve good results with reliable performance. Previous researchers have used data recorded in a lab-controlled environment to identify CI (Tanaka et al., 2017, 2016). Although their results were good, they were based upon data that does not represent real-world scenarios. Therefore, the performance of their approaches may not be reliable for the deployment environment.

This research uses in-the-wild data for people with CI, which can enlighten both computer vision and healthcare research by increasing the knowledge of the relationship between these two fields, and thus each field can benefit from the other. The datasets used in this research – IVA_{18} and the IVA_{52} – were recorded in two different environments, a clinic and a home, as described in Chapter 3. The most interesting aspect of the datasets is that the videos were recorded without any restrictions on the participants and the environment settings. The participants were not told how to sit and where to look during the session, and there was no preparation for the light conditions and the distance with respect to the camera and light. Each aspect of these challenges can affect the available techniques for facial landmark prediction.

Contribution 3: Automatic detection of ND, MCI and FMD by analysing facial cues in video recordings.

To date, research has mainly focused on detecting CI from HC using video recordings rather than detecting several health conditions from each other (Tanaka et al., 2019, 2017, 2016). In their work, ND and MCI were included in one group. This thesis contribution focuses on investigating automatic methods for differentiating several health conditions (ND, MCI and FMD) from each other using facial cues in video recordings. This task is a very challenging task even for clinicians due to the overlap between these health conditions (Wakefield et al., 2018). This is novel because this study is the first to detect ND, MCI and FMD using video recordings. The findings achieved could provide insight into the applicability of detecting these different health conditions automatically from each other. In addition, this has wide-ranging benefits for researchers and clinicians. For researchers, classifying these health conditions from each other can help to improve their understanding, facilitate new discoveries, and carry out studies that target particular conditions, resulting in more deeply focused and efficient research outcomes. This contribution can also benefit clinicians by improving the accuracy of the diagnosis, which will help clinicians with differentiating conditions with similar symptoms from each other, and providing patients with the appropriate care and treatment, resulting in much better resource management and more efficient patient care. This contribution addresses research question RQ.3 and has been published in ***ACII 2021*** (see details below). Chapters 4 and 6 present the results achieved in classifying these health conditions from each other using facial features on the IVA_{18} and IVA_{52} datasets.

Contribution 4: Developing an approach for detecting EBR that is robust to in-the-wild data.

In eye-blink detection, researchers typically use standard datasets to evaluate the performance of their approaches (Fogelton and Benesova, 2016, 2018; Pan et al., 2007). However, their approaches cannot detect eye blinks in data recorded in the wild, which consists of considerable noise (this will be discussed in Chapter 3). This research uses in-the-wild data, resulting in several challenges but not limited to: participants having variable distances with respect to the camera during the session and participants sitting in non-optimal positions, poor illumination, background noise and the appearance of other people in the camera view. As a result, these challenges make EBR detection more challenging, as it relies on calculating the eye openness ratio (eye aspect ratio, EAR), which uses the height and width of the eyes (Soukupová and Cech, 2016). An eye blink is detected when the EAR value is below a particular threshold, where the threshold is a value that determines when the eye is open or closed. Determining the threshold for datasets with healthy individuals is relatively straightforward because the participant sits relatively still in the video frames. However, this way is not sufficient for datasets recorded in the wild, such as the dataset used in this research (the IVA_{18} dataset). Therefore, an approach to detecting eye blinks using this type of recording is essential for this research because people with memory problems may show different spontaneous behaviour during a session, thus resulting in several challenges that do not just affect eye-blink detection but even face detection (Taati et al., 2019). A novel approach called the multiple thresholds (MTs) approach is proposed. This approach generates multiple thresholds for detecting blinks, resulting in having a vector of blink rates calculated corresponding to a certain range of thresholds for every participant. This approach is more robust to tackle in-the-wild data challenges.

Further evaluation was carried out to detect CI using a larger in-the-wild dataset (IVA_{52}). The experiments revealed abnormal behaviour in the classification results due to the variations in the recording environments and the devices used (e.g., laptops and smartphones) in the newly recorded data. Therefore, some improvements are made to overcome these challenges and make the MTs calculations participant-dependent instead of calculating them over all the participants' minimums and maximums. This contribution addresses research questions RQ.2 and RQ.3 and has been published in ***ACII 2021*** (see details below). The details of this contribution are explained in Chapter 4. The findings demonstrate the system's applicability for reliable CI detection in real-world settings, paving the way for deployment.

Contribution 5: Investigating the importance of the EBR and head movement features as an indicator of CI.

Investigating the automatic detection of CI using the EBR and head movement cues could confirm the work conducted in the medical literature and provide clinicians with a rapid decision in the primary diagnosis. In the medical literature, several studies found that the EBR in the people with CI group is abnormally higher than in the HC group (Albert et al., 2011; Kocagoncu et al., 2022; Ladas et al., 2014; Taylor et al., 1999; Woodruff-Pak, 2001). Head movement, particularly the head turn cue, has been shown to be an indicator of CI because patients with CI often come with a partner or caregiver to seek support for answers they cannot remember (Bouchard and Rossor, 2007; Larner, 2012). According to Fukui et al. (2011), the head turn cue indicates a CI regardless of the partner's presence.

To the best of our knowledge, this research is the first to use the EBR and head movement cues for the automatic detection of CI using the IVA_{18} and IVA_{52} datasets (see Chapter 4). The findings of this work showed that detecting CI automatically in such data using these cues is a promising area. The findings showed that ND and MCI participants have longer blinks than FMD participants. This work could help clinicians save time and effort from doing long procedures to measure this cue in order to diagnose the patient, considering the large number of people who come to the clinic and the small number of people who agree to do such sessions. This contribution addresses research question RQ.3 and has been published in ***ACII 2021*** (see details below).

The work described in Contributions 3 and this contribution was validated on a public depression dataset due to the association between CI and depression (Muliyala and Varghese, 2010), and to compare the results with related work, as explained in Chapter 7. This evaluation shows how particular features can help to provide performance comparable to related work that uses advanced techniques such as neural networks. This contribution and its validation have been published in ***FG 2023*** (see details below).

1.4 List of Publications

1. Fatimah Alzahrani, Bahman Mirheidari, Daniel Blackburn, Steve Maddock, and Heidi Christensen. "Eye Blink Rate Based Detection of Cognitive Impairment Using In-the-wild Data." In 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII).

2. Fatimah Alzahrani, Bahman Mirheidari, Daniel Blackburn, Steve Maddock, and Heidi Christensen. "Investigating Visual Features for Cognitive Impairment Detection Using In-the-wild Data." In 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG).

1.5 Thesis Structure

Figure 1.1 presents a diagram of the organisation of this thesis along with the research questions. The content of the following chapters is summarised as follows:

- **Chapter 2: Background and Related Work.** This chapter presents the nature of CI in general, its stages from normal ageing to the progression of dementia, and its effect on an individual in general and his/her non-verbal communication in particular. It also illustrates a potential new direction of clinical assessment to assess and diagnose related conditions using facial cues. It discusses different automatic techniques to detect CI that include various methods and algorithms to detect CI. This chapter also reviews the relevant literature on approaches for face detection, eye blink detection and head movement estimation. Finally, the background on machine learning algorithms and the evaluation metrics used in the literature and in this research is covered.

- **Chapter 3: In-the-wild Data.** This chapter focuses on analysing the challenges of datasets with video recordings, identifying how the in-the-wild definition is different from one dataset to another, the differences in the recordings settings and participant behaviour, and how these differences affect the processing of the videos **(RQ.1)**. It includes a description of the main datasets used in this thesis (i.e., the IVA_{18} and IVA_{52} datasets) and compares them with datasets mentioned in related work.

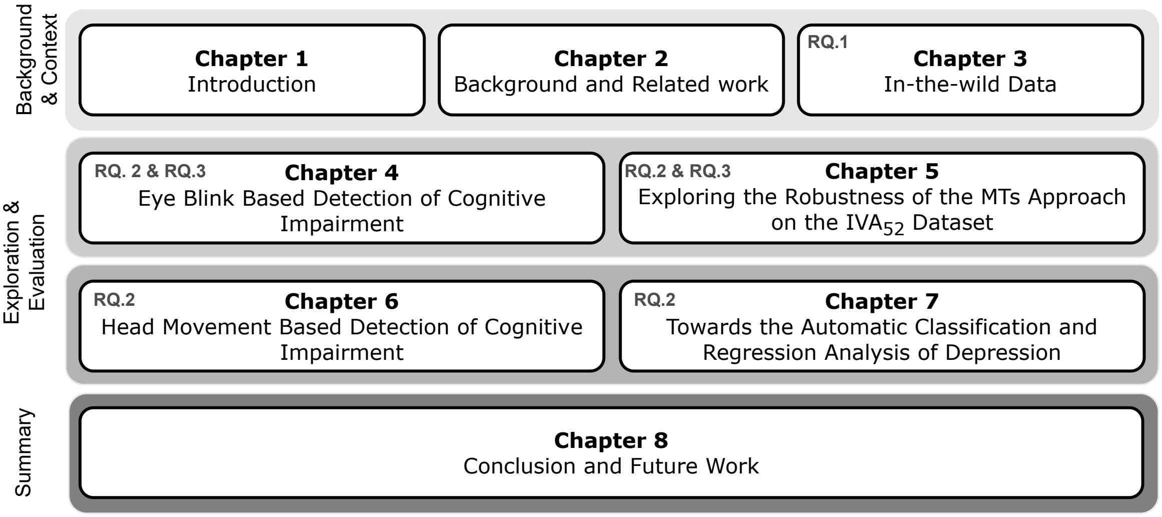

Background & Context
Exploration & Evaluation
Summary
Chapter 1
Introduction
Chapter 2
Background and Related work
RQ.1
Chapter 3
In-the-wild Data
RQ. 2 & RQ.3
Chapter 4
Eye Blink Based Detection of Cognitive Impairment
RQ.2 & RQ.3
Chapter 5
Exploring the Robustness of the MTs Approach on the IVA52 Dataset
RQ.2
Chapter 6
Head Movement Based Detection of Cognitive Impairment
RQ.2
Chapter 7
Towards the Automatic Classification and Regression Analysis of Depression
Chapter 8
Conclusion and Future Work

Fig. 1.1 Organisation of the thesis. The investigated research questions in each chapter of this thesis are indicated (RQ: research question).

- **Chapter 4: Eye Blink-Based Detection of Cognitive Impairment.** This chapter presents two approaches for the automatic detection of the EBR. The first uses a single threshold to calculate the EBR based on the participant's mean and the standard deviation. The approach is evaluated on a standard dataset collected for eye blink detection, used by many previous studies, and then applied to the IVA_{18} dataset for people with memory problems. The second approach uses the minimum and maximum values over all the participants to calculate MTs and results in multiple blink rates for each participant (**RQ.2** and **RQ.3**). The crucial advantage of the MTs approach is that it overcomes the challenges in the IVA_{18} dataset. The performance of this approach is tested on the IVA_{18} dataset and evaluated using Supervised machine learning classifiers.

- **Chapter 5: Exploring the Robustness of the MTs Approach on the IVA_{52} Dataset.** This chapter evaluates the developed MTs approach described in Chapter 4 by applying it to the IVA_{52} dataset of people with memory problems and an HC group (**RQ.2** and **RQ.3**). This evaluation led to improving the MTs approach and made it participant-based instead of taking the minimum and maximum over all the participants. It involves testing different approaches for the detection of outliers and carrying out different classification tasks. The chapter ends with a discussion of the findings.

- **Chapter 6: Head Movements Based Detection of Cognitive Impairment.** This chapter illustrates the features extraction of HTR, HTSF, HMSF and LLFs using the IVA_{18} dataset **(RQ.3)**. The performance of these features is evaluated for each feature individually, and then the fusion of all features, including the EBR feature. This chapter also evaluates the performance when feature selection is applied to choose the optimal features. Several supervised machine learning classifiers are used. This chapter also evaluates those features by applying them to the extended dataset of people with memory problems and HC group. It includes analysing the performance using machine learning classifiers. It presents, discusses, and compares the obtained performance with the IVA_{18} dataset.

- **Chapter 7: Towards the Automatic Classification and Regression Analysis of Depression.** This chapter evaluates the performance of the approaches and features explained in Chapters 4 and 6 by comparing the performance achieved with that reported in related work. The evaluation consists of two tasks: classification and regression **(RQ.3)**. The classification detects depression, and the regression predicts the severity of the depression. The chapter also presents, discusses and compares the performance obtained with that reported in previous studies.

- **Chapter 8: Conclusion and Future Work.** The final chapter includes the conclusion of the thesis and presents potential directions for future work.

Chapter 2

Background and Related Work

> *"Things not to say to someone with mental illness:*
> *Ignore it. Forget about it. Fight it. You are*
> *better than this. You are overthinking."*
> *— Nitya Prakash*

Normal ageing is a process in which the body and brain gradually change over several years. The changes range from eye trouble and hearing loss to memory loss and a decrease in agility. As people age, they can experience cognitive changes that are common and related to normal ageing, such as difficulty recalling names of people or places, taking longer to process new information, and occasionally forgetting appointments or events (Ahmad et al., 2013; Smith et al., 2022). Although the body and the brain begin to slow down in normal ageing, a person's intelligence remains relatively steady. Due to the high lack of awareness worldwide, people may mistake cognitive impairment (CI) for age-related cognitive decline. CI, however, can exhibit more frequent and disturbed symptoms. Figure 2.1 compares several clinical symptoms between normal ageing and CI.

CI can progress to a pre-clinical stage of mild (MCI) and then to dementia, which is irreversible and has no treatment. Hence, diagnosing it in its very early stage is very important. However, there is no single clinical test. Available tests are complicated and time and effort-consuming, as described in Chapter 1. This puts a heavy burden on health and economic services. An objective screening method based on physiological and behavioural cues is needed to improve the current objective screening method. Recently, research has shown significant progress in utilising affective computing and social signal processing as a diagnostic tool, especially for CI (Fei et al., 2019; Kunz et al., 2007; Tanaka et al., 2019, 2017). These methods particularly depend on face and body monitoring algorithms to capture CI-related behavioural changes. In order to investigate how facial behaviour could be used for CI detection, it is important to first understand what CI is and highlight prior work into associated facial behavioural cues to gain an understanding of how facial cues could be employed to build a cost-effective tool for CI detection.

This chapter is organised as follows. Section 2.1 provides background information regarding CI, its stages and symptoms. Section 2.2 discusses the significant role of facial behaviour in human communication and the effect of CI on people's facial behaviour. Section 2.3 reviews previous investigations concerning the detection of CI using facial cues. Section 2.4 presents a review of previous work on computer vision techniques designed to extract facial cues. Section 2.5 gives a brief description of the machine learning techniques and evaluation metrics used in this research. Section 2.6 then summarises the major limitations of previous work and the key points of this chapter.

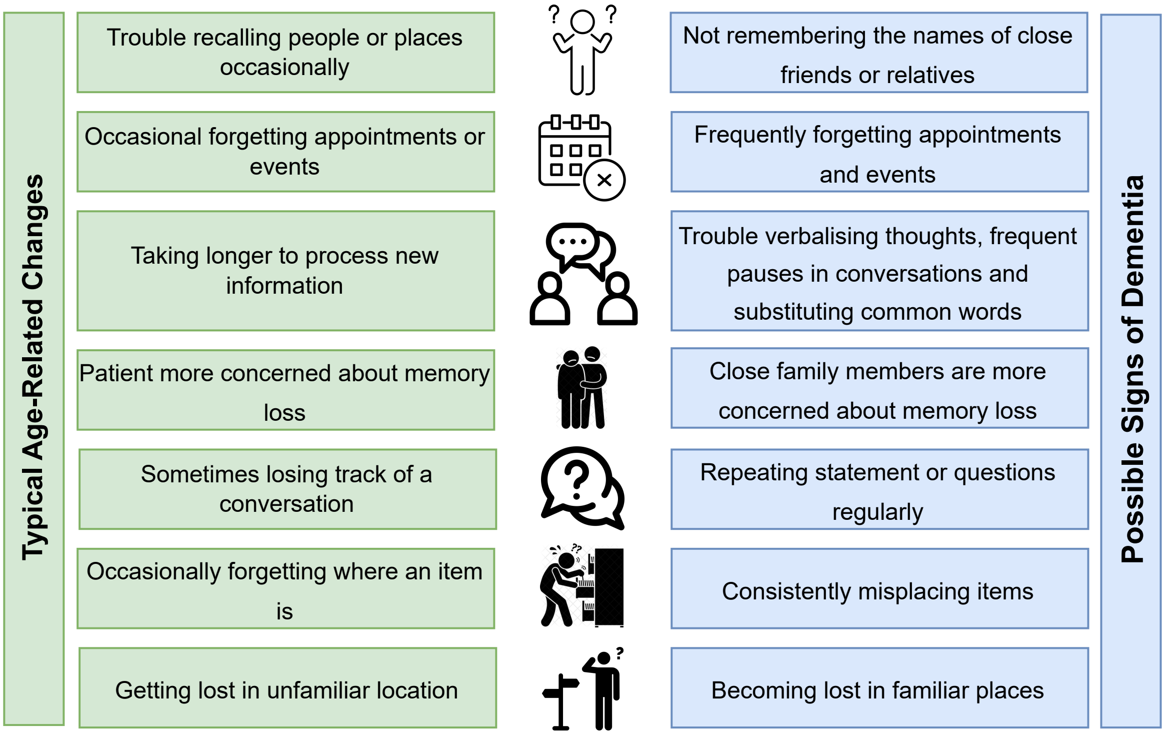

Typical Age-Related Changes
Trouble recalling people or places occasionally
Occasional forgetting appointments or events
Taking longer to process new information
Patient more concerned about memory loss
Sometimes losing track of a conversation
Occasionally forgetting where an item is
Getting lost in unfamiliar location
Not remembering the names of close friends or relatives
Frequently forgetting appointments and events
Trouble verbalising thoughts, frequent pauses in conversations and substituting common words
Close family members are more concerned about memory loss
Repeating statement or questions regularly
Consistently misplacing items
Becoming lost in familiar places
Possible Signs of Dementia

Fig. 2.1 Clinical differences between typical age-related changes and possible indicators of dementia Smith et al. (2022).

2.1 Cognitive Impairment

CI can be defined as a condition that affects several functions, such as memory, attention, changes in mood and difficulty speaking and understanding and recognising people and/or places. A wide range of potential factors has been investigated in the pathology of CI regarding sociodemographic-related factors, health and health behaviour-related factors, and cardiovascular health-related factors, as depicted in Figure 2.2, (Bornstein et al., 2014; Cermakova et al., 2020, 2017a,b; Horackova et al., 2019; Seblova et al., 2019).

Clinicians have divided the stages of developing Alzheimer's disease (AD) or other types of dementia into preclinical, MCI, and dementia (see Figure 2.3). The following will describe MCI, in which the symptoms start to appear to the doctor; dementia, in which the symptoms show significant change; and other important memory conditions that show similar symptoms to MCI and dementia.

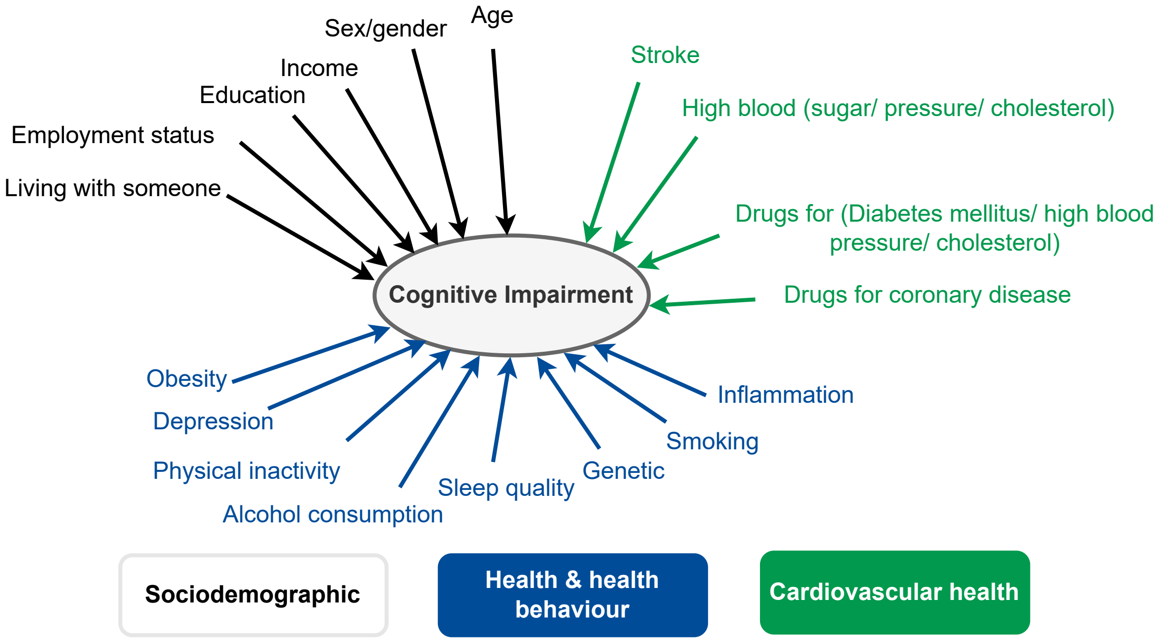


Fig. 2.2 Potential factors that may cause cognitive impairment Seblova et al. (2019).

2.1.1 Mild Cognitive Impairment

MCI is the intermediate stage between normal ageing and AD or any other type of dementia. Several causes can increase the risk of developing MCI, such as ageing and depression, as shown in Figure 2.2. MCI's symptoms can include frequently losing things, forgetting appointments or events, difficulty in coming up with words and following a conversation compared with people of the same age, and having trouble with managing finance. Whilst people with MCI can progress into dementia, some people may become stable or even improve (Petersen et al., 2018). The reason for this is unknown.

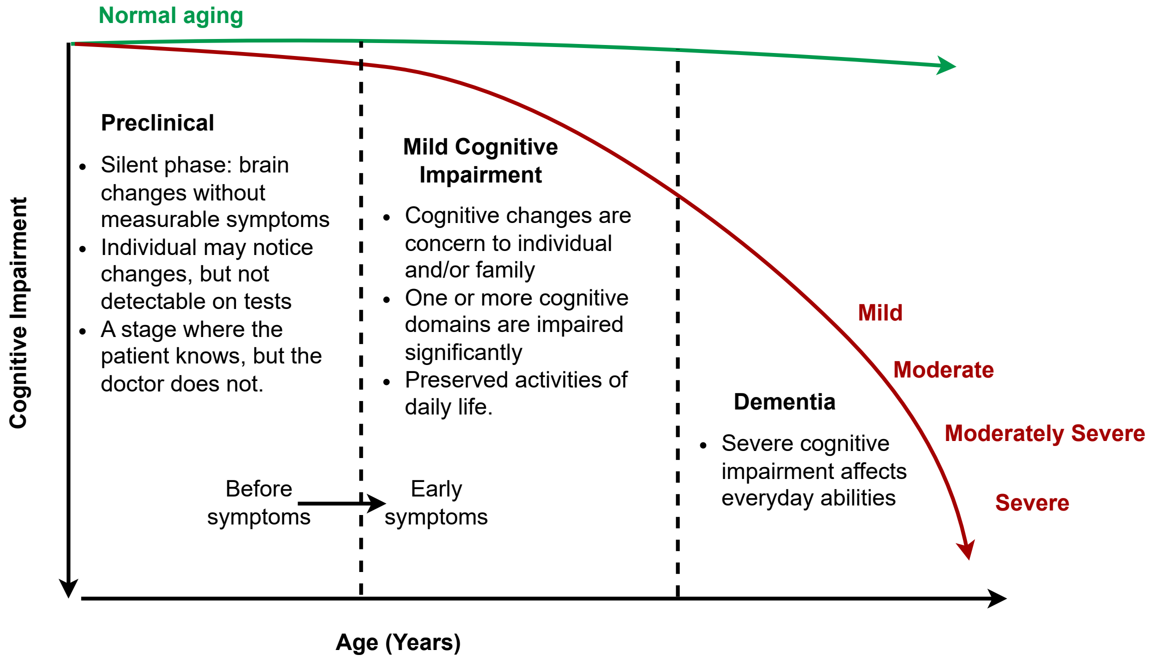

Normal aging
Cognitive Impairment
Preclinical
• Silent phase: brain changes without measurable symptoms
• Individual may notice changes, but not detectable on tests
• A stage where the patient knows, but the doctor does not.
Mild Cognitive Impairment
• Cognitive changes are concern to individual and/or family
• One or more cognitive domains are impaired significantly
• Preserved activities of daily life.
Dementia
• Severe cognitive impairment affects everyday abilities
Mild
Moderate
Moderately Severe
Severe
Before symptoms
Early symptoms
Age (Years)

Fig. 2.3 Progression of cognitive decline from normal ageing to dementia or Alzheimer's disease (Source: https://www.mind.uci.edu/dementia/mild-cognitive-impairment/).

2.1.2 Dementia

Dementia is a clinical syndrome that can significantly influence people's ability to express their feelings and complete everyday tasks. This results in unexpected behaviour and the loss of independence in daily life (Beville, 2012; Kim and Kang, 2017; Kirk and Berntsen, 2018; McKhann et al., 2011). There are several forms of dementia,such as AD, Vascular Dementia (VaD), Mixed Dementia, dementia with Lewy bodies (DLB), Frontotemporal dementia (FTD), Parkinson's disease (PD), and other (e.g., Creutzfeldt-Jakob disease, depression and multiple sclerosis), as depicted in Figure 2.4. It can be seen that the most common form of dementia is AD, with about 62% of dementia cases.

Both types of AD and FTD are brain diseases that cause irreversible dementia and can be referred to as neurodegenerative disorder (ND) (Bayles et al., 1987). There is an association between the specific effects caused by dementia and which part of the brain is harmed because each part of the brain is responsible for certain bodily functions, such as decision-making and memory (Alz.org, 2019). These functions cannot perform well if damage occurs to their specific brain areas. Most dementia types generally display similar symptoms, such as difficulties in concentrating, language loss, memory loss, difficulties in answering questions, declines in attention span, difficulties in completing sentences, depression and cognitive decline (Kavé and Goral, 2018). In terms of non-verbal behaviour cues, some types of

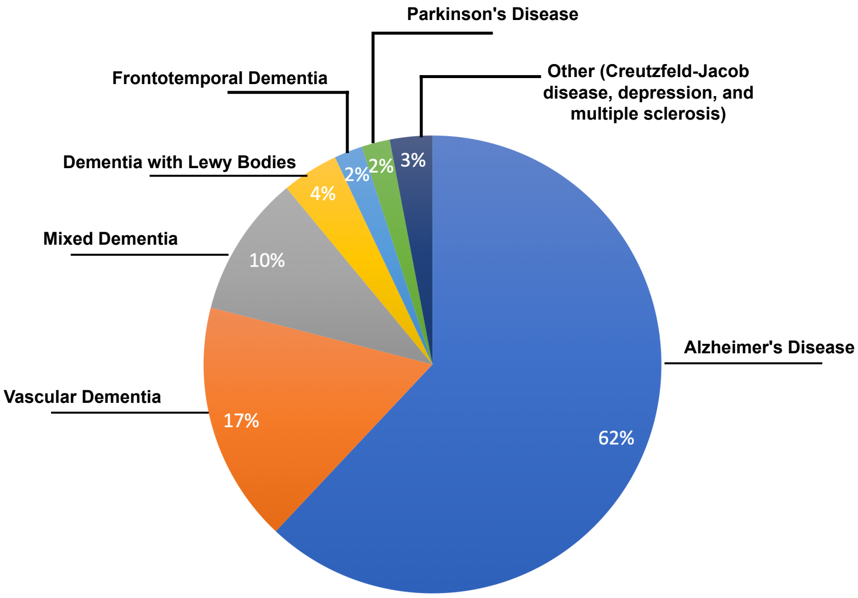


Fig. 2.4 The different forms of dementia Foggin (2018).

dementia make people more emotional (they can become very happy or sad and tearful). In addition, loss of blinking control in people with dementia was confirmed by (Ladas et al., 2014) compared with controls. Regarding memory loss, dementia significantly affects short-term rather than long-term memory (i.e. people with dementia usually forget recent conversations or events) (Knopman et al., 2003; Soysal et al., 2017). This is why dementia patients often rely on support from a partner or a caregiver (Fukui et al., 2011; Larner, 2005, 2012; Soysal et al., 2017).

2.1.3 Other Causes of Memory Complaints

Other causes of memory complaints can share the same symptoms of dementia, but these conditions are reversible, unlike ND. These conditions can cause confusion for doctors due to overlapping symptoms and can make the diagnosis very challenging. These conditions are explained as follows:

- **Functional memory disorder (FMD)** is a major cause of memory issues. It is caused by distress and distraction, which leads to difficulty in coping with a lot of information, poor concentration and attention, and storing and retrieving memory contents (Schmidtke et al., 2008). FMD can be referred to as a medical and psychological syndrome without a physical cause. This condition can be treated and improved. FMD differs from cognitive dysfunction, which includes several psychiatric diseases, such as depression, psychosis and dissociative states.

- **Depression**, which is one cause of dementia, can affect people's cognition in a way that may look like dementia symptoms. Depression can be referred to as either clinical depression or a depressive disorder that affects mood, behaviour, thinking and emotions. Those with depression lose interest in work and social activities (Association et al., 2013). The many symptoms of depression include becoming more emotional, tearful, hopeless and sad, having memory problems, sleeping and eating disturbances, a lack of energy and thoughts of suicide (Association et al., 2013). The most common symptoms in older people include memory loss, behavioural changes, decreased socialness and increased thoughts of suicide. Generally, emotions such as sadness and hopelessness can appear on a person's face. A number of facial cues, such as eye movement, facial expressions, head gestures and smile intensity, are associated with depression and can be used to differentiate depressed people from healthy ones (Cohn et al., 2009; Cummins et al., 2015; Scherer et al., 2013). Extensive research has shown that people with depression have less and slower head movement than healthy individuals (Girard et al., 2014) and exhibit reduced smile intensities and smile duration.

Even though several specialists may consider MCI similar to the early stage of dementia, MCI symptoms can be improved and stabilised over time (Petersen et al., 2018). Hence not every MCI patient develops AD, and the reason for this is unknown. Currently, early diagnosis of CI involves many steps, such as taking a person's medical history, physical examination and tests for cognitive assessment (e.g., montreal cognitive assessment and minimal mental status examination), which are conducted by an expert clinician, as explained in detail in Chapter 1, Section 1.1. These approaches are complicated, costly in terms of time and effort, and need professional neurological expertise. Clinicians have observed an intrinsic relationship between CI and facial cues, and this will be covered in detail in the next section.

2.2 Cognitive Impairment and Facial Behaviour

Diagnosing CI based on observable behavioural signals has not been fully embraced in conventional neurology. However, interest in research in this field has been growing by focusing on behavioural cues, particularly facial cues, for CI. Given the importance of early diagnosis for this condition, the need for a multi-faceted approach to finding a true objective cue is likely to be high. It is, therefore, necessary to review prior work on these facial cues, such as eye blink rate (EBR) (Ladas et al., 2014) and head movement (Larner, 2012; Soysal et al., 2017), to identify CI.

To accomplish this, this section first considers the significance of facial behaviour in human communication and how cognitive changes can have an impact on facial behaviour. Second, it presents a review of some previous studies that employed the use of these facial cues clinically for CI detection.

2.2.1 Facial Behaviour

Nonverbal behaviour is often seen as a form of human communication and a continual signal that conveys essential information about people's emotions, personalities and mental states (Heylen, 2006; Richmond et al., 2008). Eye and head movements are considered important to nonverbal behaviour in social communication and are associated with cognitive state (Jongkees and Colzato, 2016; Nakano, 2015). Both have received significant attention.

Spontaneous eye blink is an unconscious expression which is communicated through the frontal, parietal, and temporal brain regions (Mota and Lins, 2017). Simply, the term 'eye blink' is defined as a reflex that rapidly closes and opens the eyelids. An eye blink usually lasts between 100 and 400 milliseconds (Stern et al., 1984). It could be an incomplete blink when the eye is partially closed, mostly due to dry eye syndrome (Portello et al., 2013). An extended blink is when the eye closure lasts between 70 milliseconds and 1 second. Some people blink many times in succession; for example, double or even quadruple blinks are possible.

Moreover, a person's EBR plays a significant role in eye movements, fixations, emotional expressions and visual cognition (Delgado-García et al., 2002, 2003). Ageing affects the average blink rate by increasing it from about 24/minute at age 40 to 49 years to 32/minute at age 80 to 89 years (Sun et al., 1997). In addition, environment-related factors can affect the blink rate and duration, such as temperature, brightness, air conditions and relative humidity (Sun et al., 1997). Several studies have focused on studying eye blink in the context of conditions affecting cognition (King and Michels, 1957; Ponder and Kennedy, 1927).

In terms of brain activity, a relationship has been found between eye blink and cognitive state (Jongkees and Colzato, 2016; Nakano, 2015). Spontaneous eye blinks reflect cognitive states, and certain activities can cause an increase in a person's blink rate. For example, the blink rate for a person increases during speaking (in adults) (von Cramon and Schuri, 1980), conversation (Bentivoglio et al., 1997), memorising (De Jong and Merckelbach, 1990), stress, positive mood and emotions, fatigue, pain, physical activity, disease and when expressing anger or excitement (Chermahini and Hommel, 2012; De Padova et al., 2009; Ponder and

Kennedy, 1927; Sun et al., 1997); and a person's blink rate may decrease in visual tracking and reading activities (Argilés et al., 2015; De Jong and Merckelbach, 1990).

Head movement plays a crucial role in facial behaviour. It is an easy cue to understand and conveys valuable information. For instance, Maynard (1987) and Boholm and Allwood (2010) studied the role of head movement during a conversation and communicative feedback and observed that people usually tend to use particular head movements, such as nods in a vertical direction (i.e., starting from up and then down, or down and then up), tilt by leaning the head to one side, and head turn to the left or right side for expressing a positive or negative attitude, asking for a turn (Hadar et al., 1984), indicating an acceptance or rejection (Boholm and Allwood, 2010) and helping in assessing pain (Werner et al., 2018) or mental status (Larner, 2012). Such facial cues have been investigated clinically for detecting CI and early signs of dementia in the next section.

2.2.2 Facial Cues and Cognitive Impairment

In general, people with dementia's facial expressions and emotions show significant changes since their expressive abilities are maintained until the severe stages of the condition (Lee et al., 2013, 2022). Studies have observed increased negative facial expressions, even in mild AD (Heilman and Nadeau, 2022; Smith, 1995), and increased lip movements, blinking, closed eyes and opening lips while in pain (Asplund et al., 1991; Jonell et al., 2021). Aside from spontaneous facial expression research, expressions have also been studied during painful medical procedures such as injections (Kunz et al., 2007). This research focuses on two facial cues, EBR and head movement, which are explained as follows.

Eye blink Rate

Previous work has considered EBR as a reliable and easy biomarker to use for measuring the brain's central dopamine activity (DA) (Mackert et al., 1991; Taylor et al., 1999). DA is associated with the early stages of MCI, which indicates that EBR is related to the early stages of MCI. To explain in more detail, it is important to first understand the relationship between dopamine and the early stages of MCI. Braak and Braak (1997) and Braak et al. (2005) observed that neurofibrillary pathology in the very early stages of AD is initially found in the entorhinal cortex and then develops in the area of the hippocampus and later in the rest of the limbic system and other cortical areas, as shown in Figure 2.5. Due to the importance of the entorhinal cortex and hippocampus in the episodic memory (Burianova et al., 2010; Woodard et al., 2009), it is expected that episodic memory is impaired in the very early stage of ND (i.e., the pre-clinical stage of MCI) (Kocagoncu et al., 2022). Dopamine is an important neurotransmitter in the hippocampus and the limbic system, so it is not surprising that people with MCI can show a deficiency in DA, as reported previously (Albert et al., 2011). From that, it is suggested that dopamine is directly associated with a number of cognitive functions, such as episodic memory, learning and executive functions (Kok, 2022), which constitute neuropsychological predictors of transformation from MCI to AD (Arnáiz and Almkvist, 2003). Thus, changes in the dopamine system in AD may play a significant role in the gradual cognitive deterioration seen in pathological ageing (Martorana et al., 2010).

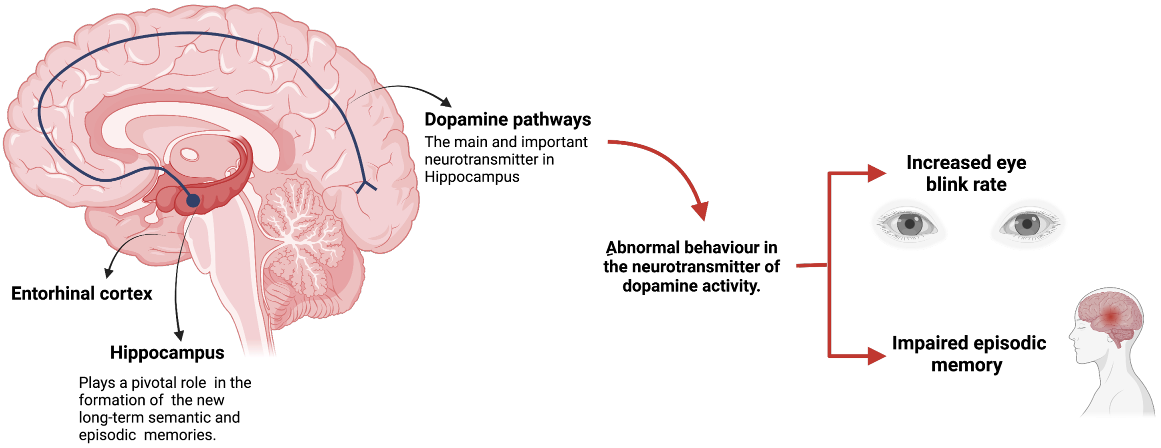

Dopamine pathways
The main and important neurotransmitter in Hippocampus
Entorhinal cortex
Hippocampus
Plays a pivotal role in the formation of the new long-term semantic and episodic memories.
Abnormal behaviour in the neurotransmitter of dopamine activity.
Increased eye blink rate
Impaired episodic memory

Fig. 2.5 Some parts of the limbic system play a significant role in episodic memory, including the entorhinal cortex, hippocampus and dopamine neurotransmitter. Any abnormal behaviour in the dopamine activity can lead to impaired memory and an increase in the eye blink rate (EBR).

Ladas et al. (2014) employed the EBR to examine DA in MCI and figure out how DA is associated with cognitive performance. They reported that MCI had considerably greater EBR than healthy controls (HC). Their work is considered to have been the first in measuring the EBR for the MCI group. On the other hand, relatively little research has looked into the validity of EBR as a biological cue of DA in healthy elderly people (De Padova et al., 2009; Goschke and Bolte, 2014; Sun et al., 1997). Ladas et al. (2014) also conducted a correlation study with previous work in terms of the average age of their participants and their average EBR. Importantly, previous studies involved participants with an average age of 65.22 years (SD=11.2), which is quite similar to the average age of 67.52 years of the HC in (Ladas et al., 2014)'s study. Regarding the EBR, the HC in previous work had an average EBR of 20.27 blinks/minute, which is nearly equal to the average EBR of 20.24 blinks/minute of the HC group in (Ladas et al., 2014). Consequently, it could be reliably concluded that the EBR in the MCI group was abnormally high rather than the contrary (the healthy group had a lower

EBR than expected). This conclusion suggests an increase of DA in the central nervous system. This finding of increased DA in the MCI group contradicts prior research that found a decreased DA with ageing (Volkow et al., 1998) and in AD (Kemppainen et al., 2003). However, this contradiction in findings may be consistent with the theory that an overall neurotransmitter imbalance in the pre-clinical stage of MCI leads to a high blink rate for the MCI group.

Head Movement

A number of clinicians have found that people having CI turn their heads frequently to their accompanying person when asked a question (Bouchard and Rossor, 2007). In addition, NHS Evidence Clinical Knowledge Summary reported that CI should be diagnosed "if, when you ask the person a simple question, they immediately turn to their partner in the so-called head-turning sign" (Bouchard and Rossor, 2007). Accordingly, several studies have explored the prevalence and benefits of the head turn cue in diagnosis by conducting prospective observational research of day-to-day clinical practice in a memory disorders clinic. Accordingly, a number of cross-sectional studies suggested an association between CI and head movement, in particular, the head turn (Fukui et al., 2011; Larner, 2005, 2012; Soysal et al., 2017). Soysal et al. (2017) applied a geriatric assessment test, which is a multidimensional test used to assess the functional ability, cognition, mental health, physical health, social life and environment of a patient (Elsawy and Higgins, 2011). The researchers assessed the head turn and attended with (AW) cues in addition to a geriatric assessment test conducted on 529 patients with a mean age of 75.67 years. They found that people who brought a caretaker/partner with them (+AW) had lower scores on all their tests, such as the minimal mental status examination (MMSE), the Montreal cognitive assessment (MoCA) and the cognitive state test (Babacan-Yildiz et al., 2013). Moreover, the results showed that the head turn cue is affected by the partner's presence, and both could be used as indicators of CI.

In addition, Larner (2018) evaluated five signs mentioned in the previous work (Bonello and Larner, 2016; Ghadiri-Sani and Larner, 2013; Larner, 2012, 2014b; Soysal et al., 2017) to assess cognitive status: attended alone (AA), AW, head turn, applause [1] and la Maladie du petit papier [2]. The presence of CI was indicated by three signs: AW, applause and head turn. However, the other two signs (AA and la Maladie du petit papier) were considered signs of the absence of CI. The five signs were examined to compute three different metrics: the

[1] This is a neurological test which is also called the 'Clapping test' because the participants are asked to clap their hands three times as fast as possible similar to the examiner's demonstration (Bonello and Larner, 2016)

[2] Asking the patient to write his/her symptoms on paper or an iPad during the assessment in the clinics.

number needed to diagnose (NND), the number needed to predict (NNP) and the number needed to misdiagnose (NNM). The results proved that the NND and the NNP could be useful for detecting CI, but there was a risk of misdiagnosis. These findings were consistent with those of previous studies (Durães et al., 2018; Larner, 2012) that found the head turn and AW were significant cues to identify the CI and AA as a cue of the absence of CI. Although the partner's presence could affect the head turn cue, other factors, such as gender and age, may have an effect on the partner's presence.

Factors found to influence the head turn and AW cues have been explored in several studies, such as the type of dementia, gender and age (Fukui et al., 2011; Larner, 2014a; Lövheim et al., 2009). Fukui et al. (2011), for example, investigated the incidence and severity of the head turn cue in patients with AD-related disease (AD and MCI) and AD-non-related disease (DLB, VaD and progressive supranuclear palsy(PSP)). They found the severity of the head turn cue was significantly higher in the AD-related group, specifically the female gender significantly contributed to the incidence and severity of the head turn cue. Fukui et al. (2011) suggested that a possible explanation for their findings is that women at heart feel easier to depending on someone else to face difficulties with them. In contrast, men feel obligated to handle difficulties without help. Their explanation is consistent with previous work that explored the differences between women and men with CI regarding the prevalence of behavioural symptoms and found that depression and 'help-seeking' is more frequent in women than men. However, regressive and aggressive behaviour is more frequent in men than women (Lövheim et al., 2009). Furthermore, Larner (2014a) mentioned that age can play a significant role in the presence of the partner.

On the other hand, those findings are contrary to those of previous studies, which suggested that gender and age are independent of the partner's presence (Holland and Larner, 2013; Larner, 2005). In addition, the head turn is a cue of CI regardless of the partner's presence (Fukui et al., 2011). The possible explanation for this is the different groups' division because Fukui et al. (2011) considered only AD and MCI as one group and other dementia types in another group, whereas Soysal et al. (2017) included all dementia types in one group with MCI and AD. To date, the effect of a cultural factor on these factors and the head turn cue has received scant attention in the research literature.

Diagnosis in clinics is considered costly in terms of money and time for the patients and the clinics due to the high number of people who go to the clinic and the lack of clinicians to interview and assess all of them. In addition, some people do not feel comfortable going to the clinic and talking about their problems to a doctor. However, the facial cues discussed above could have a wide potential for use as objective cues. They could be used as a low-cost and low-effort tool for automatically capturing and analysing information from a large population and providing instant feedback and advice for patients (Scherer et al., 2013). Ultimately, investigating objective facial cues could help to improve the accuracy of clinicians' diagnoses and reduce the socio-economic cost associated with this condition (Costanza et al., 2014; JH Balsters et al., 2012).

2.3 Automatic Detection of Cognitive Impairment Using Facial Cues

Numerous studies have attempted to automatically identify CI by employing computer vision techniques with machine learning. However, few studies into CI detection have focused on facial cues (e.g., eye movement, smile expression and facial action units) to demonstrate the potential of their approaches (Barral et al., 2020; Fraser et al., 2019; Tanaka et al., 2017, 2016). This section reviews some studies that have explored some facial cues in CI detection.

2.3.1 Eye Movement

Eye movement changes can be very subtle, making detecting them difficult during standard clinical assessment. Recently, however, they have attracted interest from researchers as a valuable cue for detecting the CI (Beltrán et al., 2018; Endo et al., 2017). A range of approaches have been used to detect CI, such as advanced eye-tracking technology (for a review, see (Beltrán et al., 2018)) and computer vision techniques that detect the face and then localise the iris during a video recording (Endo et al., 2017).

Eye movement tracking involves measuring saccadic and anti-saccadic eye movements and gaze fixation because they hold rich information (Anderson and MacAskill, 2013). Saccadic eye movement can be defined as a quick movement of the eye (taking 30–80 ms to complete) and is gauged by looking at an object that appears in various locations on a screen (Holmqvist et al., 2011). In contrast, anti-saccadic eye movement is gauged by the cessation of gazing at a particular object and instead looking in the opposite direction. Gaze fixation is maintaining the eyes' concentration on a particular spot accurately. Gaze metrics can be detected while reading, viewing pictures or movies, or performing other cognitive tasks that require eye-hand coordination and visual memory recognition (Endo et al., 2017; Lagun et al., 2011; Oyama et al., 2019). It is now well established from a variety of studies that specific eye movement patterns can convey valuable information to distinguish people with CI (Boucart et al., 2014; Pereira et al., 2014), including increased staring, increased blinking and increased fixation instability, which means the difficulty of keeping the gaze on a specific location (Coubard, 2016).

Furthermore, many eye movement problems in people with dementia indicate cognition deficiencies, such as visuospatial ability, attention, memory, inhibitory control and executive functioning (Coubard, 2016; Pereira et al., 2014). A primary factor influencing this impairment in eye gaze for people with dementia is a lack of inhibitory control. To examine inhibitory control deficits, different saccadic and anti-saccade tasks (Crawford et al., 2005; Lagun et al., 2011) have been used to detect early deterioration during MCI and dementia. People with dementia find some tasks very difficult, such as directing eye gaze towards a stimulus and moving voluntary eye gaze away from a stimulus (Wilcockson et al., 2019).

2.3.2 Facial Expressions

Although some studies have been carried out on the automatic detection of CI using eye movement, there have been just a few empirical investigations into other facial cues, such as facial expressions and facial action units (FAUs), carried out by one Japanese research group. Tanaka et al. (2016) proposed an approach to detecting CI from HC using data recorded by an intelligent virtual agent (IVA). The dataset comprised 18 participants from whom the researchers extracted audio-visual features (smile as a facial expression). Although the smile feature did not show any significance, they reported a good performance in classifying people with CI from HC with 94% accuracy using a support vector machine (SVM). However, when Tanaka et al. (2017) increased the sample size to 29, they found that the smile feature was significant and obtained an efficient performance with an accuracy of 93% using SVM. They found that people with CI smiled more than HC. A possible explanation for this finding could be that when people smile out of frustration, they are having trouble answering questions (Hoque and Picard, 2011). One limitation is that their approach cannot be applied to non-Japanese people because the researchers used a facial model based on Japanese women to compare the facial expressions of the participants.

Subsequently, Tanaka et al. (2019) investigated the use of FAUs, introduced by Ekman et al. (2002), extracted using the OpenFace toolkit to detect dementia on data with recordings from 24 participants. Their findings showed that lip activity, FAUs and eye gaze were significant features, with an accuracy of 82%. However, the data used was recorded in a lab-controlled environment, which is expensive, requires a lot of effort, and does not represent real-life scenarios.

2.4 Techniques for Extracting Facial Cues

The previous section reviewed related work in terms of the use of facial cues for dementia detection. This section provides an analysis of several techniques that can be used for extracting facial cues: detection of eye blink and head movement estimation. Approaches for eye blink detection include motion-based approaches, template matching-based approaches, facial landmarks-based approaches and other approaches. For head movement estimation, four approaches have been described: 2D appearance-based approaches, regression-based approaches, model-based 3D head registration approaches and deep learning-based approaches.

2.4.1 Eye Blink Detection

Eye blink detection is a major area of interest in various fields, such as e-learning, gaming, assistive technologies, security, detecting fatigue and health care applications (Al-Rahayfeh and Faezipour, 2013a; Azim et al., 2014; Dong and Wu, 2005; Li et al., 2018a; Li and Feng, 2019). A number of techniques have been developed for eye blink detection using computer-vision-based approaches (Bacivarov et al., 2008; Fogelton and Benesova, 2018; Luo et al., 2019). Two phases usually need to be considered to detect an eye blink: detecting the position of the eyes and tracking them through the video frames recorded with a webcam (Al-Rahayfeh and Faezipour, 2013a). Such datasets contain several challenges to accurate eye tracking, which must be considered, such as determining the eye size and openness, as described in Chapter 3.

This research focuses on prior work used on data collected with webcams. These algorithms could help to pave the way for application on common devices (e.g., laptops and smartphones) to decrease complexity, cost, and effort and increase simplicity, accessibility and usability. Several approaches developed to detect blink are reviewed in the following section based on eye optical flow motion, eye template matching and facial landmarks. Table 2.1 compares the performance of previous studies for eye blink detection using different datasets. The table is organised according to the type of approach used. Their results generally are good, but the reason for this could be the nature of the datasets used to assess their approaches, which were captured in a lab-controlled environment.

Motion-based Approaches

Several studies have attempted to detect eye blink based on eyelid motion. Motion vectors usually calculate the changes from one frame to the next. Divjak and Bischof (2009) employed eyelid motion for blink detection, which comprises several steps. First, the locations of the participant's face and the left and right eye were detected by three different classifiers (Rosten and Drummond, 2006) approach. Second, they were tracked using a Lucas-Kanade tracker (Tomasi and Kanade, 1991). When the person's face is not frontal, the classifier cannot detect it. Then, the optical flow motion for the face region was calculated and normalised. Finally, adaptive thresholding was used to detect the blink. The authors followed the study of (Heishman and Duric, 2007) to detect a blink with some modifications in the calculation of the optical flow. The adopted approach comprised several steps: 1) the optical flow was calculated for the face region, 2) because eyelid motion incorporates head motions, compensation was performed based on the previously extracted head movement, 3) normalising the optical flow due to the size of different faces, 4) the direction of the dominant vertical eye movement was estimated and 5) adaptive thresholding of the processed flow data was used to detect blinks. The approach was assessed on their data, which included ten videos for three participants, recorded using a monitor-mounted webcam in an office environment where the participant faced the camera. They achieved an average accuracy of up to 91%. For general evaluation, their approach achieved 97% on the ZJU data (Pan et al., 2007), which consisted of 80 videos for 20 participants.

Mohanakrishnan et al. (2013) explored motion vectors to detect blinks by calculating the average motion vector and the similarity between each vector within the eye region. This resulted in a threshold used to classify blinks with 97% accuracy on their dataset, including 4052 images and 203 blinks. Drutarovsky and Fogelton (2014) extracted six motion vectors from the eye region and calculated the average motion and variance to feed to a state machine (SM). There was an SM for each eye to determine whether the eye closure was a true blink or not. Unlike Divjak and Bischof (2009), Drutarovsky and Fogelton (2014) considered the challenge of small head movements in their approach.

Fogelton and Benesova (2016) developed the proposed approach by (Drutarovsky and Fogelton, 2014) to overcome the problem of the variance in eye region size and head movement using the motion vector-based approach, which was calculated by the Gunnar-Farneback tracker in the eye region. The motion vectors were normalised by interocular distance to resolve the differences in the eye region size. Then, normalised motion vectors were fed to the SM with standard deviation and time constraints. An SM was used to identify whether eye closure was a genuine blink or not. Their approach outperformed related work on the ZJU dataset and the Eyeblink8 (Drutarovsky and Fogelton, 2014) dataset, which included 8 videos, with F1-measure of 99% and 93.3%, respectively.

Fogelton and Benesova (2018) introduced a method for blink completeness detection. First, they computed the motion vectors for the eye region using the dense optical flow approach following (Farnebäck, 2003). Second, the vertical and horizontal components were employed with the time difference between the consecutive frames due to their significance in the SM based on the previous work (Fogelton and Benesova, 2016). Then, they adopted a bidirectional RNN architecture (Schuster and Paliwal, 1997) to classify each frame into non-blink (0), complete blink (1) and incomplete blink (2). They reported the best results on the majority of the public datasets (i.e., ZJU, Eyeblink8, Basler5 and Research nights) compared to their previous work (Fogelton and Benesova, 2016).

Template Matching-based Approaches

Another method for determining eye closure and eye blink is to use an open eye template (Chau and Betke, 2005; Grauman et al., 2001, 2003; Królak and Strumiłło, 2012). Grauman et al. (2001) detected the eyes by calculating the correlation coefficient over time. When the correlation coefficient falls under a predefined threshold, re-initializing is triggered. Then, the blink is detected based on the changes in the correlation coefficient between the actual eye and the open-eye template and the actual eye and the closed-eye template. Finally, the correlation coefficient is binarized into open and closed eyes. The focus of their work was on people with disabilities.

Radlak and Smolka (2013) further developed the Weighted Gradient Descriptor (WGD) (Radlak and Smolka, 2012), which calculates spatio-temporal derivatives per each pixel in the eye region over time. The resulting vectors were averaged between consecutive frames into positive and negative based on location (up or down). The two vectors were weighted and used to calculate the vertical distance (the y-coordinates of those vectors) between their origin points, which is called the waveform and is used as the input signal. Opening and closing the eye were represented by positive and negative signal peaks. Then, the signal was filtered, and a cut-off point with zero was used to identify the local minimum and maximum peaks, representing the detected blinks. To assess their approach, they introduced a dataset of five people using a Basler 100 fps camera, referred to as the Basler5 dataset (Radlak and Smolka, 2012). They achieved up to 90% accuracy and reported 98.8% accuracy on ZJU. In the evaluation, they only used the right eye of the participants.

Later, Malik and Smolka (2014) utilised a template matching approach that was computed based on the Local Binary Patterns (LBP) technique. LBP is a descriptor used to capture features of the eye region. First, an initial process was conducted to build an open-eye template using several images where the eye was open and not moving. Then, this template was used to compare the histogram of the LBP of the subsequent frames. After the signal was filtered, the sharp peaks between the template and the histogram of the current frame were considered to be detected blinks. The efficiency of their approach was measured on two datasets, the ZJU and Basler5 datasets, with detection rates of 99% and 94.2%, respectively.

Bhowmick and Mustafa (2021) and Kamanga and Lyimo (2022) also employed the open-eye template matching approach and achieved an efficient performance. Even though Bhowmick and Mustafa (2021) claimed that they used a dataset in real-world scenarios, they used the Haar Cascade Classifier for face detection and the Camshift algorithm for face tracking, which cannot handle challenging datasets, indicating that their dataset was not very challenging.

Facial Landmarks-based Approaches

A number of researchers have investigated the use of facial landmarks to facilitate eye blink detection (Alghowinem et al., 2013a; Bacivarov et al., 2008; Soukupová and Cech, 2016). Bacivarov et al. (2008) proposed a statistical active appearance model (AAM), which is a semi-automatic approach to detecting and tracking eye blink. This model was trained on a manually selected image with different variations in the eye states (e.g., open, closed, partially open). These images were annotated manually with 74 points in the eye region before sending them to the eye AAM model. Then, the model could be used to detect and track the eyes. They evaluated their model on two datasets: the Georgia Tech Face Database [3], which included 750 images for 50 participants with 15 images for each participant and the VidTIMIT Video dataset (Sanderson and Paliwal, 2004), which included 43 participants, and reported accuracies of 96.87% and 100%, respectively.

Eye aspect ratio (EAR) is another approach for blink detection. Soukupová and Cech (2016), developed an algorithm that worked in real-time applications with a standard camera to detect the EBR. The algorithm was conducted by finding the eye region landmarks and computing the EAR, which was then used in the SVM to determine whether the eye is open or closed. Their algorithms performed well in tests with two different datasets. The simplicity of their algorithm was clear, but there were some limitations in their work. First, making the blink duration fixed for every participant as people are different in blink duration, especially people with health conditions. Moreover, the head orientation was not considered, and neither was the background noise. Finally, the study was completely dependent on annotated data.

Maior et al. (2020) adopted the EAR approach introduced by (Soukupová and Cech, 2016) to detect the eye blink by using 15 consecutive EARs for the machine learning model to return one category (i.e., open eye, short blink or long). Three models were used to evaluate these three categories: multilayer perception, random forest and SVM. They recorded 282 samples for 13 participants, including 109 for open eyes, 95 for short blinks, and 78 for long blinks. They found that SVM gave the highest performance with an accuracy of 94.9%.

Other research teams, such as Navastara et al. (2020) and Utaminingrum et al. (2021), also employed the EAR principle for eye blink detection. Utaminingrum et al. (2021) evaluated

[3] ftp://ftp.ee.gatech.edu/pub/users/hayes/facedb/

their work on 42 images for ten different participants and obtained an accuracy of 90.5%. However, Navastara et al. (2020) investigated the use of EAR and the uniform local binary pattern as combined features with SVM in the blink detection, and they achieved an accuracy of 95.5% on a dataset collection from (Song et al., 2014). Dewi et al. (2022) also employed the EAR approach following Soukupová and Cech (2016), but with some modifications. They introduced a new threshold based on the EAR with the name Modified EAR, which involved using the interquartile range to calculate the lower bound as a threshold.

Other Approaches

Several studies have attempted to explore other techniques. For example, Wang et al. (2017) used the Contour Circle algorithm and Adaboost approach together to detect eye blink for fatigue recognition. Their approach was evaluated on their dataset of five people recorded for one minute in a laboratory environment with a blink detection accuracy of 96.6%. A neural network is another approach to identifying the eye blink (de la Cruz et al., 2022; de Lima Medeiros et al., 2022; Nanthini et al., 2022). Li et al. (2018c) proposed an approach that used both convolutional neural networks (CNNs) and recurrent neural networks for eye blink detection from a video to differentiate between the real talking face and the generated one. They compared their performance to the EAR approach of (Soukupová and Cech, 2016) and reported a higher obtained performance of 99% on their dataset compared to 79% using EAR. Even though the performance was higher by a significant margin, they used Dlib for face detection and landmarks prediction. This indicates that the dataset used is not challenging because Dlib can only detect frontal and semi-frontal faces.

de Lima Medeiros et al. (2022) used two models– CNNs and SVM– and introduced two new datasets: the youtube eye-state classification (YEC) and the autonomus blink dataset (ABD). Both models were trained using the YEC dataset and were evaluated on several datasets, such as ZJU, Eyeblink8, talking face and ABD. Moreover, Gawande and Badotra (2022) employed hybrid methods for eye blink detection using CNNs to accurately localise the eye area and then used the EAR to determine the eye state (i.e., closed or open). Finally, the blink is identified based on the calculated correlation coefficient, which is classified as short and long blinks.

Table 2.1 Comparison of performance for eye blink detection approaches when applied to lab-controlled datasets (Note: the metrics used for performance are different.)

Paper	**Dataset**	**Reported F-measure (or as otherwise noted)**
Motion-based Approach		
Divjak and Bischof (2009)	Their dataset ZJU	Accuracy=91.0% 97.0%
Mohanakrishnan et al. (2013)	Their dataset	Accuracy=96.9%
Drutarovsky and Fogelton (2014)	ZJU Eyeblink8 Talking face	99.8% 99.9% 99.8%
Fogelton and Benesova (2016)	ZJU Eyeblink8	99.0% 93.3%
Fogelton and Benesova (2018)	ZJU Eyeblink8 Talking face RN test set Basler5	97.6% 91.0% 97.1% 87.9% 94.5%
Template Matching-based Approach		
Grauman et al. (2001)	8 participants	Accuracy=95.6%
Chau and Betke (2005)	(Grauman et al., 2001) dataset	Accuracy=95.3%
Radlak and Smolka (2012)	Basler5	Accuracy=76.1%

Radlak and Smolka (2013)	Basler5
	ZJU
		Accuracy=83.4%
		Accuracy=98.8%

Malik and Smolka (2014)	Basler5
	ZJU
		Accuracy=94.2%
		Accuracy=99.2%

Bhowmick and Mustafa (2021)	50 participants	Accuracy=97.3%
Kamanga and Lyimo (2022)	ZJU	Accuracy=96.5%

Table 2.1 Continuation of Table 2.1

Facial Landmarks-based Approach		
Bacivarov et al. (2008)	Georgia Tech Face Database VidTIMIT	96.9% 100%
Maior et al. (2020)	13 participants	94.9%
Dewi et al. (2022)	talking face Eyeblink8	95.0% Precision=99.0%
Other Approaches		
Wang et al. (2017)	5 participants	96.6%
Li et al. (2018b)	CEW dataset (Song et al., 2014)	99.0%
de Lima Medeiros et al. (2022)	ZJU Eyeblink8 Talking face ABD	92.3% 86.9% 95.0% 92.6%
Gawande and Badotra (2022)	(Alhakeem et al., 2020) dataset	92.0%

2.4.2 Head Movement Estimation

A considerable amount of literature has been published on automatic head movement estimation (Abate et al., 2022; Khan et al., 2021). These studies have used different techniques, making it difficult to organise them in a particular classification framework due to the overlapping between these techniques. The following section describes different techniques for estimating head movement using 2D facial image properties and 3D face pose.

2D Appearance-based Approaches

In these approaches, there is an assumption about the strong relationship between a 3D face pose and the 2D facial image properties. This relation is described by using many images and techniques of statistical learning to train a model. Then, visual features are extracted from the statistical distribution of the training images. The resulting model is used to differentiate between different head poses or movements. This technique has been used in the past and has shown very good results in (Burl and Perona, 1996; Firintepe et al., 2020; Jebara, 1995; Ma et al., 2015; Qin et al., 2022). Some drawbacks could affect the performance of this approach and result in errors, such as head movement under extreme differences in illumination, facial expressions, occlusions, participants with eyeglasses and facial hair.

Regression-based Approaches

These techniques use functional mapping from image space to different head poses. It requires a set of labelled images for the training, and then a new model can be created to discriminate head poses for new data (discrete or continuous). The main problem with these techniques is that the regression tool may not learn a proper mapping due to the challenge of the high dimensionality of the image. This issue can be resolved by using principal component analysis or linear discriminant analysis to reduce image dimensionality and then feed them to support vector regression (SVR) with efficient performance (Li et al., 2000, 2004). Previous studies used the localised gradient histograms with SVR and showed a better performance (Murphy-Chutorian et al., 2007). Regression-based approaches can only be used on low-dimensionality features. As a non-linear regression tool, neural networks were also utilised in (Brown and Tian, 2002; Duda et al., 2006; Little et al., 2005). Although neural networks are very efficient and straightforward to implement and update, the performance drops significantly if the images are not annotated correctly.

Model-based 3D Head Registration Approaches

These techniques detect some points from a 2D image and project them onto a 3D face model. For example, Alghowinem et al. (2013b) used the previously mentioned AAM (Bacivarov et al., 2008) with pose from orthography and scaling with iterations (DeMenthon and Davis, 1995) approach to detect at least four points in a 2D image and then project them onto a 3D model and then calculate the orientation and translation of the 3D model. Meyer et al. (2015) estimated the head pose using measured depth data, which were registered to a morphable model. Another study used a 3D morphable model and online 3D reconstruction to estimate the full head pose (Yu et al., 2017). Similar work has been done by fitting only the face rather than using the full head because of the low quality of depth data (Ghiass et al., 2015).

Deep Learning-based Approaches

Deep Learning approaches (DL), particularly CNNs, have achieved the highest performance compared to feature-based machine learning approaches. Several issues and limitations of the traditional machine learning approach were resolved by moving to DL (Baltrušaitis et al., 2012; Baltrusaitis et al., 2018; Kuhnke and Ostermann, 2019; Ranjan et al., 2017).

For example, (Baltrušaitis et al., 2012) proposed a 3D Constrained Local Model (CLM-Z) to use depth and intensity information to detect facial features and track them across video sequences. The depth information helped to mitigate the effect of poor illumination and when there was no intensity in the signal due to lighting conditions. More recent work by Baltrusaitis et al. (2018) outperformed the previous state-of-the-art techniques for extracting facial landmarks, and head pose estimation. They used a Convolutional Experts Constrained Local Model (CE-CLM) (Zadeh et al., 2017) for facial landmarks detection and tracking. In the internal CE-CLM structure, a 3D representation of the facial landmarks was calculated and then projected onto the image using an orthographic camera. Once the facial landmarks had been detected, this helped to accurately estimate the different head poses by solving the *n* point in the perspective approach (Hesch and Roumeliotis, 2011).

In reviewing the techniques mentioned above, in general, the performance of their approaches gave good results. However, their approaches were evaluated on data recorded in a lab-controlled environment for eye blink detection and on data recorded with some challenges for head movement estimation. Those datasets will be described in more detail in Chapter 3. Therefore, it is possible that their results do not reflect reliable performance. In this thesis, for eye blink detection, the EAR approach (Soukupová and Cech, 2016) was adopted due to the development of the facial landmark-based approaches in the field of computer vision for such a task. Regarding head movement estimation, the deep learning-based approach was used in this thesis due to the challenges in the dataset used, which will be described in the following chapter.

2.5 Machine Learning Related Background

Machine learning is categorised as a sub-discipline of artificial intelligence. It aims to automatically enhance the performance of the computer algorithms developed for a specific task. There are several types of machine learning approaches, such as supervised learning, semi-supervised learning, unsupervised learning, reinforcement learning and deep learning (Samuel, 1959). In *supervised machine learning*, the models are learned by an example. It provides the machine learning with the input data X with output Y, which can be referred to as training data, and returns a mapping function $f_x : X \longrightarrow Y$ that presents the relation between X and Y. Then, the mapping function is used to classify the test data, which are unseen samples. Support vector machine (SVM), k-nearest neighbours (KNN), logistic regression (LR), and decision trees (DT) are examples of such approaches (Bishop and Nasrabadi, 2006). An example of supervised learning is a disease diagnostic system.

In *unsupervised learning,* machine learning takes a collection of unlabelled data X and studies the correlations and the relationships of the data in order to create a model that can take input data X, and either transforms it into another vector or value to solve a problem. Two types fall under the umbrella of unsupervised learning, which are clustering and dimensionality reduction. In clustering, the data is grouped into similar sets, and the model returns the id of the cluster for each feature vector in data X. K-means and Gaussian mixture model are examples of clustering (see Chapter 6). In dimensionality reduction, the feature vector is reduced to fewer features than the input data X. An example of the dimensionality reduction approach is the Recursive Feature Elimination Cross-Validation (RFECV) (see Chapter 6). *Reinforcement Learning* in machine learning depends on rewards and errors to learn from its environment. It is provided with a feature vector of a state as input and outputs an action to execute in that particular state. It does this by exploring different possibilities and evaluating each result to find the optimal action. Reinforcement learning is designed for particular problems where decision-making is sequential, such as gaming, playing, driving a car and robotics.

Deep learning is a sub-field of machine learning that uses multi-layer structures of neural networks to build 'an artificial neural network' that replicates the human brain and can learn and make its own decisions. There are different types of deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). These algorithms are used for various tasks, and each one of them has its unique characteristics and applications. CNNs algorithms are primarily developed for tasks involving image processing, object detection and facial recognition. RNNs are algorithms that are designed specifically for tasks involving processing sequential data. It is suitable for tasks involving natural language processing, text generation, language translation and speech recognition.

The selection of the appropriate machine learning classifiers for this thesis depends on both the data size and the objectives of this work. Following similar successful healthcare research (Alghowinem et al., 2013a; Mirheidari et al., 2017; Tanaka et al., 2019, 2017), this thesis employs SVM, KNN, LR and DT classifiers. These choices are based on the fact that they usually do not require a sufficiently large amount of training data to work well, whereas deep learning models do. The following explains the chosen classifiers in more detail.

2.5.1 Support Vector Machine

The SVM is a supervised machine learning algorithm, which is commonly used for different classification tasks involving healthcare applications (Cummins et al., 2015; Tanaka et al., 2017). When used in a regression problem, it is defined as support vector regression (SVR). The goal of SVM is to find the optimal hyperplane in an N-dimensional space to classify the data points. An optimal hyperplane is a plane with a maximum margin from both classes. The hyperplanes are decision-makers that help to differentiate the data points, and their dimensions depend on the number of features. For instance, if the feature dimension is 2, the hyperplane is just a line. However, if the number of features is 3, the hyperplane will be a 2-dimensional plane. The hyperplane with maximum margin is used to classify the test data. The data points that are close to the maximum margin are called support vectors. These data points affect the position and the orientation of the hyperplane, which means that removing them will change the hyperplane position. Different kernels could be used with SVM, which differ in their mapping approach for the data points. The common ones are linear, RBF and polynomial kernels (Bishop and Nasrabadi, 2006). In the case of a non-linear kernel, the data points are mapped into high dimensional space to be classified.

2.5.2 K-Nearest Neighbours

The KNN is a discriminative machine learning algorithm, which depends on a simple idea of similarity. It is one of the most efficient supervised machine-learning algorithms that can be used for classification and regression problems. The main technique is based on measuring the distance between the data points using the Euclidean distance function because it is widely known and used as a distance metric (Bishop and Nasrabadi, 2006; Murphy, 2012). The distance values are sorted in ascending order. Then, the algorithm searches for the top k values from the sorted distances that are close to the test data. Thus, these test data points are assigned to that class. For a classification problem, majority voting is considered. However, for the regression problem, the average of the label values is taken instead of the majority.

2.5.3 Logistic Regression

LR is a predictive algorithm based on the probability concept. It makes predictions of the probability of an event occurring and then feeds these predictions into a sigmoid function to map every prediction into a probability value between 0 and 1 (Murphy, 2012). The types of LR are binary and multi-linear functions. The classification is done based on setting a decision boundary (threshold). For instance, assume that there are two classes, A and B, and that the decision of the data instance belongs to class A if the probability of this instance is above a threshold value, which is 0.5. If the returned value is 0.7, the data instance is classified as class A. However, if the value is 0.3, the observed instance is classified as class B. The parameters of the LR are estimated from the training data using the maximum likelihood estimation (MLE). This MLE is calculated using an optimisation algorithm, such as gradient descent (Murphy, 2012). L1 and L2 regularisation strategies in LR are widely used and important to avoid overfitting. That is to say, regularisation could help to reduce the error of generalising and provide better performance on an unseen dataset. The advantage of LR is its simplicity, ease of implementation, and speed at classifying test data.

2.5.4 Decision Trees

A DT is a straightforward machine learning technique, easy to implement and interpret. It is represented as a tree-like model of decisions, which involves splitting the data recursively into smaller partitions based on various rules applied at the node level. This algorithm is non-parametric and does not need any assumptions about space distribution. The cost of the accuracy is calculated in each split, and the split with the lowest cost is chosen. That is why it is called the 'greedy algorithm'. The cost function tries to determine the branches (groups) with identical responses. The optimal MLE is calculated using the greedy optimisation algorithm to find the optimal modal of the tree. When there are many features, the number of splits will be large, resulting in a huge tree and then overfitting. Therefore, the pruning approach reduces the overfitting and the tree complexity by removing the branches with low-importance features. DT can perform feature selection implicitly. However, the DT classifier could show instability in cases, where small changes in the data could create a completely different tree (Bishop and Nasrabadi, 2006).

2.5.5 Performance Evaluation

To measure the performance of those classifiers mentioned above, some metrics are used. Thus, this section mainly reviews the metrics used for performance evaluation in data related to healthcare. All the mentioned metrics, which are discussed, were used in this thesis.

Accuracy, Precision, Recall and F-measure

The most popular performance metrics for healthcare applications are accuracy, precision, recall and F-measure. These metrics are calculated based on the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN), which are defined as follows:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{2.1}$$

$$Precision = \frac{TP}{(TP+FP)} \tag{2.2}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{2.3}$$

$$\text{F-measure} = 2 \cdot \frac{Precision \cdot Recall}{(Precision + Recall)} \tag{2.4}$$

The accuracy measures the proportion of TP and TN in a whole population. Precision measures the percentage of the TP from the predicted positive instances. However, recall measures the percentage of actual positive that is labelled as positive. The F-measure can be defined as the mean of its precision and recall and could be used when the data classes are uneven. If there is an imbalanced data problem, unweighted accuracy, recall, precision and F-measure are used (Fahad et al., 2021; Gupta et al., 2020). Unweighted accuracy is significant because it offers each class equal weight. It is also not affected by minority classes. The unweighted accuracy is known as the average equal accuracy of individual classes. The individual class accuracy is calculated by the number of samples correctly predicted divided by that class's total number of samples. However, weighted accuracy is known as overall accuracy and is calculated by the number of samples correctly predicted divided by the total number of samples.

Confusion Matrix

To define the performance of a classification model, a confusion matrix is used, as shown in Table 2.2. The confusion matrix provides a summary and a visualisation of classification results. The percentages of correct and incorrect test instances are summarised for each class, which is the key to the confusion matrix. The confusion matrix consists of TP, TN, FP and FN (Gan, 2020).

Table 2.2 A confusion matrix.

		Predicted	
		Positive	Negative
Actual	Positive	true positive (TP)	false negative (FN)
	Negative	false positive (FP)	true negative (TN)

Mean Absolute Error and Root Mean Squared Error

The most common metrics for summarising and assessing the quality of the machine learning model include mean absolute error (MAE) and root mean squared error (RMSE). The MAE measures the error by subtracting the predicted value from the actual one, takes the absolute value for the error value and then calculates the mean for all the error values. The equation of MAE is presented in Equation 2.5, where n is the sample size, f_i is the predicted value and y_i is the actual one.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{2.5}$$

The RMSE can be defined as the standard deviation of the errors. It calculates the squared values of the difference between the predicted value and the actual one, takes the average of the squared values, and then the root of the value is considered, as shown in Equation 2.6.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2} \tag{2.6}$$

2.6 Summary

This chapter has reviewed previous work regarding the definition of CI, its potential causes, and the stages of the progression from normal ageing and MCI to dementia. Then, the effect of CI on people's facial cues was described, and the vital role of these facial cues was highlighted by discussing the factors that could have an effect on them. An insight into the potential of automatic CI detection using facial cues was presented. In addition, techniques for extracting facial cues were discussed. Then, the traditional supervised machine learning classifiers and evaluation metrics utilised in previous work and in this thesis were explained.

Most studies in detecting CI have been carried out by only one Japanese research group, using a few facial cues, such as eye movement, smile expression, eye gaze and FAUs. However, those studies were limited in the number of participants and their usability in only a particular country. A search of the medical literature revealed some studies which found that EBR and head movement, specifically the head turn rate, are valuable cues to identify CI. Very little is currently known about the effect of other facial cues in the automatic detection of CI, and the use of different head movements has not been clinically or automatically investigated.

The research to date has tended to focus on using a lab-controlled environment with limited challenges included in the data capture process (e.g., poor illumination and slight head movement) rather than using data recorded in the wild. Their approaches' performance has achieved very good results. However, when in-the-wild data is used, the performance may dramatically decline because they do not consider the significant challenges and the new variabilities that affect the performance and make the approach unreliable, especially for healthcare applications. Therefore, using data that represents real-life scenarios is very important to ensure reliable performance. According to the literature review, such data is called 'in-the-wild data'. There is no agreed definition of this term, and Chapter 3 consequently provides a definition of the term 'in-the-wild data' and presents more information about the differences between the datasets used in previous work and a comparison with the dataset used in this research.

The rest of this thesis will focus then on investigating these facial cues in detecting CI using in-the-wild data for people with CI and related health conditions. In addition, this investigation is designed to explore the association between CI and these facial cues. Whether there are factors that can play a vital role in the presence of these cues or not will also be discussed.

Chapter 3

Analysing the Challenges of In-the-wild Data

"The advice I'd give to somebody that's silently struggling is, you don't have to live that way. You don't have to struggle in silence. You can be un-silent. You can live well with a mental health condition, as long as you open up to somebody about it, because it's really important you share your experience with people so that you can get the help that you need."
— Demi Lovato

3.1 Introduction

The previous chapter provided background information regarding cognitive impairment (CI), other memory complaint causes that share similar symptoms, and their current clinical and automatic detection approaches. The techniques for extracting facial cues and the machine learning approaches used for relevant work were also described. Previous studies have evaluated their approaches on datasets recorded in a lab-controlled environment and achieved good results. However, the performance of their approaches may not be reliable because of the lack of in-the-wild data for evaluating state-of-the-art techniques and for the automatic detection of health conditions. In addition, most researchers used professional cameras to record the data rather than common consumer devices (e.g., laptop webcams and smartphones). Such professional cameras are not available in every home, are not affordable and require expertise to use.

This chapter is structured into six sections. Section 3.2 provides a review of the datasets used for evaluating state-of-the-art techniques (e.g., face detection, facial landmarks detection, eye blink detection and head movement estimation) and the automatic detection of health conditions. These datasets' respective data collection procedures and the challenges that could affect their performance will also be discussed in Section 3.3. Section 3.4 describes how previous work used the term 'in-the-wild data' and the kind of challenges and diversity that may be involved. Section 3.5 briefly describes the data used for the research presented in this thesis, the associated challenges and the reason for considering it as in-the-wild data. In addition, a comparison will be made between the research dataset used and previous work regarding the challenges. Finally, section 3.6 presents the conclusions of the chapter's key points.

3.2 Review and Categorisation of Datasets

Extensive research has shown that having a reliable approach or model depends on the dataset used for evaluation. The assessment of an approach developed for a particular task usually depends on each dataset's variabilities. This section will describe 1) the commonly used datasets for evaluating various computer vision tasks, such as face detection, face tracking, eye blink detection, and head movement estimation, and 2) datasets specifically used for people with health conditions.

3.2.1 Commonly Used Data

This section reviews datasets used to evaluate approaches developed for specific tasks: face detection and tracking, eye blink detection and head movement estimation, which were previously discussed in Chapter 2. Appendix A provide a brief description of each dataset, including its purpose, population if indicated by the authors, the number of images or videos, challenges, limitations and availability.

Face Detection and Tracking

Table 3.1 summarises information on each face detection dataset, including population, the number of images, the source of mages and if it is indicated as in-the-wild according to the authors. Several datasets, Labelled Faces in the Wild (LFW) (Huang et al., 2008), Helen (Le et al., 2012), Labelled Face Parts in the Wild (LFPW) (Belhumeur et al., 2013) and Menpo (Zafeiriou et al., 2017), have been used for training purposes, not for testing. However, only 300 faces in the Wild (300-W) (Sagonas et al., 2013a), Annotated Faces in the Wild (AFW) (Zhu and Ramanan, 2012) and the IJB-FL dataset (Kim et al., 2016), which is a subset of the IARPA Janus Benchmark (IJB-A) (Klare et al., 2015), datasets have been used for evaluation. The IJB-FL and Menpo have extreme variation in the face pose. However, the other datasets show small face pose variations, as shown in Figure 3.1. A sample of images from these datasets is presented in Figure 3.2.

Table 3.1 Summary of commonly used data (frame-based) for face detection. A check-mark is used as an indication of whether the dataset is in the wild based on the authors' claim (W: In-the-wild).

Data	**Images**	**Source**	**W**
Helen	2330	Internet (Flickr)	✓
LFPW	3,000	Internet (Flickr, Google, Yahoo)	✓
LFW	13,233	Internet	✓
300-W	600	Internet (Google)	✓
IJB-FL	180	Internet	✓
Menpo	14,845	Internet	✓

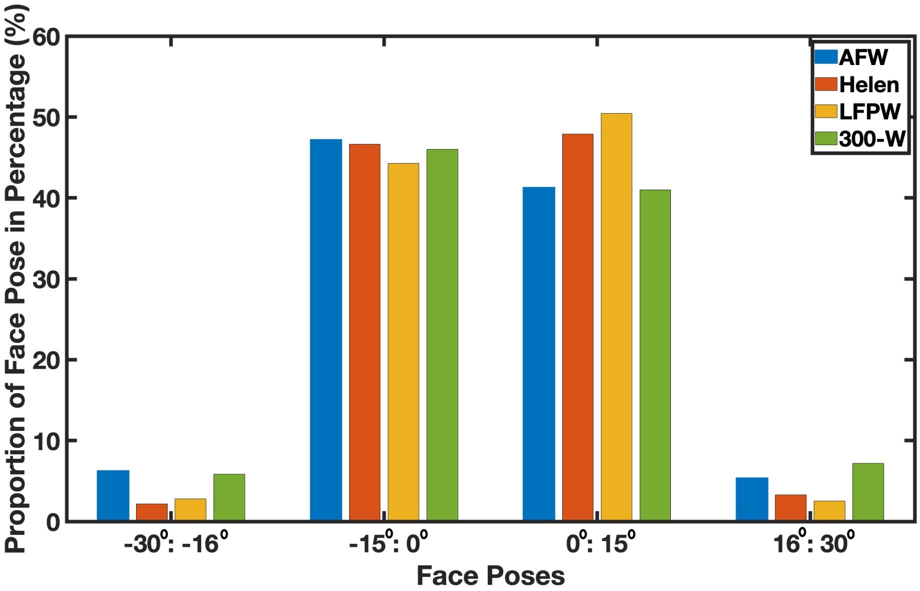


Fig. 3.1 The variations in the proportion of face pose of different in-the-wild data for face detection purposes. The proportion of face poses is gathered from previous work Sagonas et al. (2013a).

A considerable amount of literature has been published in the face alignment field. These studies have shown rapid progress in enhancing the accuracy of landmarks detection and

300-W Dataset (Helen, AFW, LFPW)

Menpo Dataset

Fig. 3.2 Sample of images from datasets for face detection.

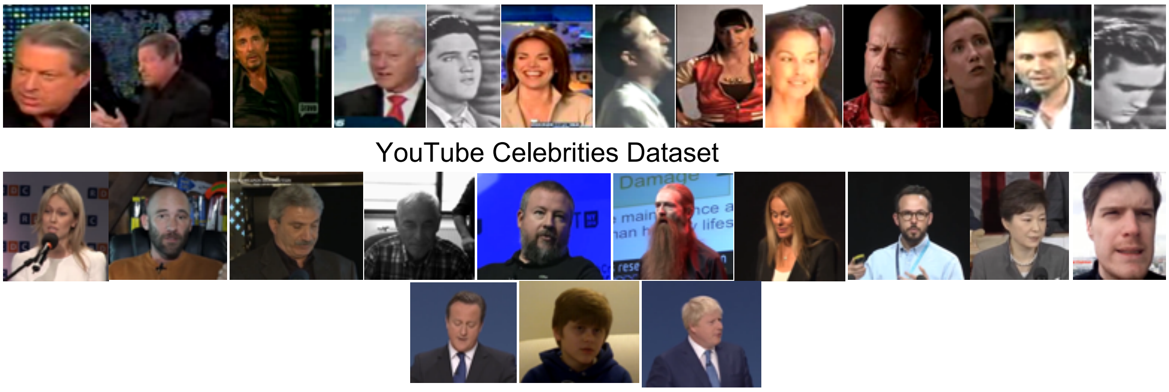
algorithm speed. This development has been possible because of the availability of the larger and in-the-wild datasets mentioned above (e.g., LFPW, Helen, AFLW, AFW, 300-W, IJB-FL, and Menpo). This motivated other investigators to examine the performance of their approaches to face tracking instead of using still images. Due to the lack of datasets for facial landmarks tracking, different challenging datasets have been collected, such as Rochester/UCSD-facial action coding system (RU-FACS) (Bartlett et al., 2006), YouTube celebrities (Kim et al., 2008), 300 videos in the wild (300VW) (Shen et al., 2015), distracted driver face (DDF) (Xiong and De la Torre, 2015) and naturalistic driving study (NDS) [1]. Table 3.2 summarises the information on each dataset used for face tracking, including population, the number of images, length if it is a video, resolution, source of the videos and if it is in-the-wild based on the authors. Figure 3.3 presents images from some of these datasets. Appendix A gives a brief description of each one of them.

Measuring the performance of face detectors on in-the-wild datasets provides a useful account of how the provided face detector model is good. However, a comparison between these datasets regarding the different face detectors used and the performance obtained cannot be conducted here for several reasons. The first reason is that different metrics are used, such as root mean square error (Xiong and De la Torre, 2015) and normalised point-to-point error (Baltrusaitis et al., 2018; Sagonas et al., 2014; Xiao et al., 2015; Yang et al., 2015).

[1] https://insight.shrp2nds.us/

Table 3.2 Summary of commonly used data (video-based) for face tracking. A check-mark is used as an indication of whether the dataset is in the wild based on the authors' claim (W: In-the-wild).

Data	Population	Videos	Length	Resolution (FPS)	Source/Camera	W
RU-FACS	29	29	-	-	-	-
YouTube Celebrities	47	1910	Avg < 3 secs	-	YouTube	✓
300-VW	-	300	Avg=64 secs	- (30)	YouTube and SEMAINE database	✓
DDF	15	15	Avg=1 min	-	-	-
NDS	-	20	Avg=64 secs	360 x 240 (15)	-	-



YouTube Celebrities Dataset

300-VW Dataset

Fig. 3.3 Images from datasets for face tracking.

Although the metric used for measuring the error is calculated based on each video and presented in a graph, the error score is not calculated by taking the mean from all the videos. Some researchers reported the mean error for particular videos that perform better than previous work (Sagonas et al., 2014). In addition, some studies have conducted qualitative facial landmarks tracking results of particular frames (Xiong and De la Torre, 2015) instead of calculating the error margin.

Eye blink Detection

To evaluate approaches for eye blink detection, a number of widely used datasets have been established. This section will describe the datasets ZJU (Pan et al., 2007), Talking face [2], Eyeblink8 (Drutarovsky and Fogelton, 2014), Basler5 (Radlak and Smolka, 2012) and

[2] $http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html$

Researcher's Night (Fogelton and Benesova, 2016). The purpose of collecting these datasets was solely eye blink detection, except for Talking face, which is built for facial landmarks detection. They can be requested from https://www.blinkingmatters.com/research. None of them is claimed to be *'in-the-wild'*. The details of each dataset are summarised in Table 3.3, and a sample of images from these datasets are shown in Figure 3.4. In addition, different published papers that have reported on the face detectors used and the results obtained from different datasets are shown in Table 3.4. Generally, their approaches achieved good results, taking into account that these datasets are not in the wild. It can be seen that previous work used face detectors that cannot handle in-the-wild cases.

Table 3.3 Summary of commonly used datasets for eye blink detection. A check-mark is used as an indication of whether the dataset is in the wild based on the authors' claim (W: In-the-wild, FSP: frame per second, RN: Researcher's night, Avg: average, min:minutes, and secs:seconds).

Data	Population	Videos	Length	Resolution (FPS)	Camera	W
ZJU	20	80	Avg = 5 secs	320 x 240 (30)	LogitechPro5000	-
Talking face	1	1	2:46 mins	720 x 576 (25)	-	-
Eyeblink8	4	8	Avg = 5 mins	640 x 480 (30)	Logitech C905	-
Basler5	5	-	-	640 x 480 (100)	Basler	-
RN 15	38	38	-	640 x 480 (15)	-	-
RN 30	69	69	-	640 x 480 (30)	-	-

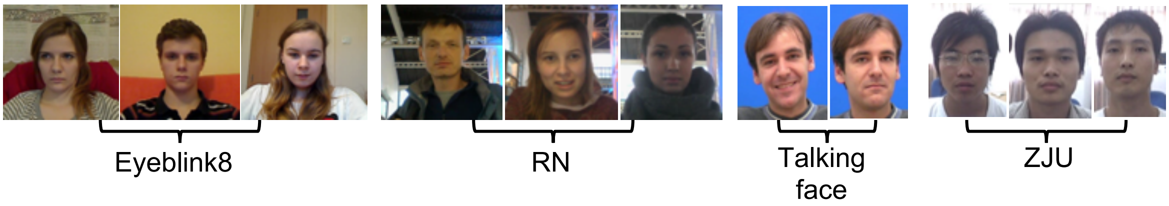


Fig. 3.4 Images from datasets for eye blink detection.

Head Movement Estimation

In addition, one study mentioned in the literature review has reviewed several studies that had attempted to estimate head movement by evaluating their work on public datasets that evolved in recent years to incorporate the complexity of environmental conditions (He et al., 2022). A comprehensive list of these datasets is listed in Table 3.5, which are BU (La Cascia

Table 3.4 Comparison between different face detector approaches and the performance of the eye blink detection obtained grouped by data (SIM: selected images manually and RN: Researcher's night, SVM: support vector machine, LBP: local binary pattern).

Data	Paper	Videos	Face Detector	Performance
ZJU	(Divjak and Bischof, 2009) (Song et al., 2014)	80 SIM	Lucas-Kanade algorithm Viola and Jones algorithm	Accuracy=97% 97%
	(Soukupová and Cech, 2016)	80	Intraface algorithm	F1=95%
	(Eddine et al., 2018)	SIM	LBP+SVM	95%
	(Fogelton and Benesova, 2016)	80	CLandmark+ Viola–Jones algorithm	F1=93.3%
	(Zhao et al., 2018)	SIM	Viola–Jones algorithm	97%
	(Fogelton and Benesova, 2018)	80	CLandmark+ Viola–Jones algorithm	F1=97%
Talking face	(Divjak and Bischof, 2009)	1	Lucas-Kanade algorithm	Accuracy=88%
	(Soukupová and Cech, 2016)	1	Intraface algorithm	F1=95%
	(Fogelton and Benesova, 2016)	1	CLandmark+ Viola–Jones algorithm	F1=94%
	(Fogelton and Benesova, 2018)	1	CLandmark+ Viola–Jones algorithm	F1=97%
Eyeblink8	(Fogelton and Benesova, 2016)	8	CLandmark+ Viola–Jones algorithm	F1=93%
	(Soukupová and Cech, 2016)	8	Intraface algorithm	F1=95%
	(Fogelton and Benesova, 2018)	8	CLandmark+ Viola–Jones algorithm	F1=97%
Basler5	(Radlak and Smolka, 2012)	5	Eye area approximation	76%
	(Radlak and Smolka, 2013)	5	Eye area approximation	83%
	(Malik and Smolka, 2014)	5	LBP	94%
	(Fogelton and Benesova, 2016)	5	CLandmark+ Viola–Jones algorithm	F1=94%
RN	(Fogelton and Benesova, 2016)	107	CLandmark+ Viola–Jones algorithm	F1=80%
	(Fogelton and Benesova, 2018)	107	CLandmark+ Viola–Jones algorithm	F1=87%

et al., 2000), Annotated Facial Landmarks in the Wild (AFLW) (Koestinger et al., 2011), ICT-3DHP (Baltrušaitis et al., 2012), and BIWI (Fanelli et al., 2013). These datasets are collected for head movement tracking, and a sample of images is presented in Figure 3.5. Appendix A gives an overview of each. As mentioned above, the performance of the head movement estimation models significantly depends on the quality of a dataset. Several published papers reported the face detectors used, and the results obtained from different datasets are set out in Table 3.6. In general, most studies achieved good results using these datasets. However, their performance declined when applied to AFLW dataset because it is in the wild.

Table 3.5 Summary of commonly used data for head movement estimation. A check-mark is used as an indication of whether the dataset is in the wild based on the authors' claim (W: In-the-wild, FSP: frame per second and secs:seconds).

Data	**Population**	**Images**	**Length**	**Resolution (FPS)**	**Camera**	**W**
AFLW	25,993	21,997	-	-	Internet	✓
BU	5	200	7 secs	320 x 240 (30)	Magnetic Tracker + Sony Handy-cam	-
ICT-3DHP	10	1400	-	-	Magnetic Tracker + Kinect	-
BIWI	20	15,000	-	-	Magnetic Tracker + Kinect	-

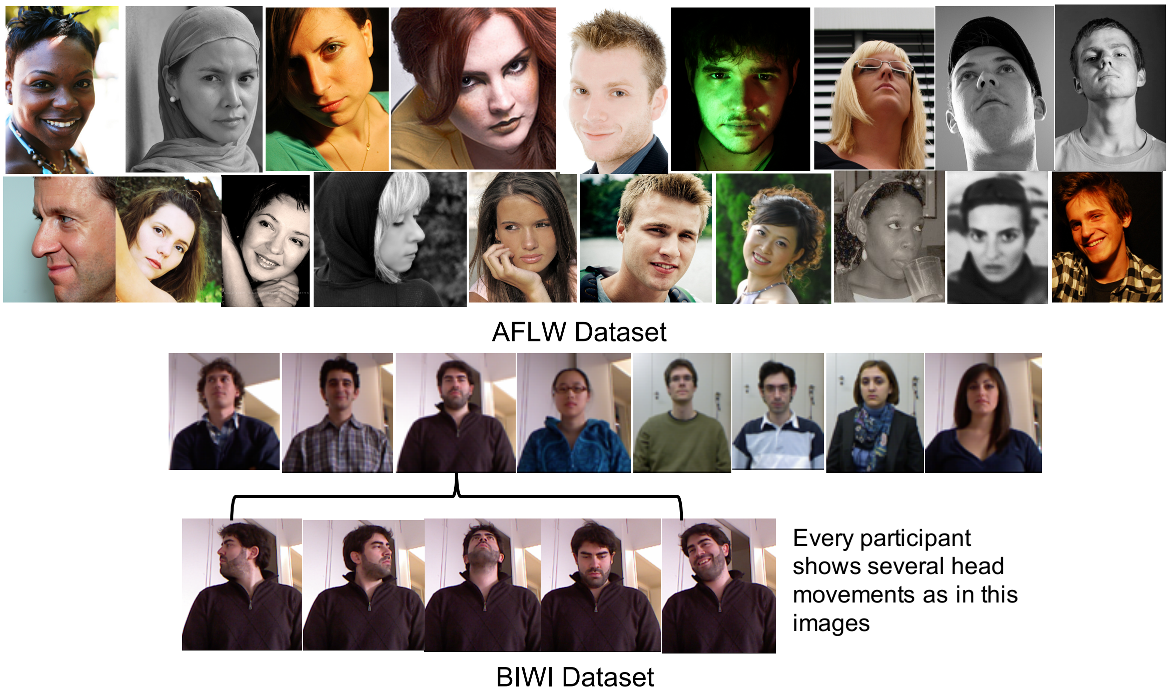

AFLW Dataset
Every participant shows several head movements as in this images
BIWI Dataset

Fig. 3.5 Sample of images from datasets for head movement estimation.

Table 3.6 Comparison between different face detector approaches grouped by data (Y: yaw, P: pitch, R: roll, 3DMM: 3D Morphable Models, and RM: Regression Model).

Data	Paper	Face Detector	Y-MAE	P-MAE	R-MAE
AFLW	(Xia et al., 2022)	DL	5.15	4.49	2.15
	(Rahmaniar et al., 2022)	DL	5.10	4.51	2.59
	(Chen et al., 2021)	DL	5.49	23.81	17.26
	(Barra et al., 2020)	RM	3.11	4.82	2.25
	(Khan et al., 2019)	DL	4.25	4.89	3.20
	(Hsu et al., 2018)	DL	3.93	4.31	2.59
	(Ranjan et al., 2017)	DL	4.71	4.14	3.93
BU	(Kazemi and Sullivan, 2014)	Dlib	23.1	13.6	10.5
	(Gou et al., 2022)	3DMM	4.80	4.60	3.00
	(Gou and Ji, 2020)	RT	5.10	4.30	3.20
	(Khan et al., 2019)	DL	2.10	2.90	2.20
	(Baltrusaitis et al., 2018)	DL	2.40	3.20	2.40
	(Baltrušaitis et al., 2016)	DL	2.80	3.30	2.30
	(Wu et al., 2017)	RM	4.90	5.30	3.10
ICT-3DHP	(Li et al., 2021)	3DMM	2.70	2.60	2.10
	(Madrigal and Lerasle, 2020)	3DMM	4.19	3.88	4.33
	(Khan et al., 2019)	DL	2.60	3.20	2.70
	(Baltrusaitis et al., 2018)	DL	3.50	3.10	3.10
	(Baltrušaitis et al., 2016)	DL	3.60	3.60	3.60
BIWI	(Chen et al., 2023)	DL	4.74	4.50	2.55
	(Celestino et al., 2023)	DL	3.81	3.78	2.73
	(Rahmaniar et al., 2022)	DL	3.27	3.19	2.43
	(Barra et al., 2020)	RM	6.21	3.95	4.16
	(Yu et al., 2017)	3DMM	2.50	1.50	2.00
	(Hsu et al., 2018)	DL	4.01	5.49	2.93
	(Kuhnke and Ostermann, 2019)	DL	4.11	4.51	3.78

3.2.2 Health Conditions Data

In terms of datasets for people with health conditions, several datasets with video recordings are used by researchers. As was mentioned in the previous chapter, depression is another condition that shares similar symptoms to CI. Also, a validation experiment will be conducted using the public depression dataset described in Chapter 7. This section, therefore, reviews a total of 13 datasets, eight for depression and five for CI, which are BlackDog (Alghowinem et al., 2012), Oregon Research Institute (ORI) (Maddage et al., 2009), AVEC2013 (Valstar et al., 2013), AVEC2014 (Valstar et al., 2014), DAIC-WoZ (Gratch et al., 2014), CHI-MEI (Huang et al., 2016), Pittsburgh (Dibeklioğlu et al., 2017), BD (Çiftçi et al., 2018), Osaka University dataset 2016 for dementia (OU2016) (Tanaka et al., 2016), OU2017 (Tanaka et al., 2017), OU2019 (Tanaka et al., 2019), IVA_{18} and IVA_{52}. Only five of the eight for depression are publicly available (AVEC2013, AVEC2014, DAIC-WoZ, Pittsburgh and BD). These datasets have been widely adopted in the reviewed studies for the detection of depression. There are also privately released datasets for depression detection. However, there are no available datasets with videos for CI for public use. The IVA_{18} and IVA_{52} datasets will be described in detail in Section 3.5. Table 3.7 provides a summary of these datasets, including the number of recruited participants, the type of assessment, video resolution, availability and if it is in-the-wild data. Appendix A gives further details of these datasets. As stated earlier, the focus is on the most popular ones.

Table 3.8 summarises the reviewed approaches, the face detectors used and the performance obtained for depression detection for the most widely used datasets. It can be seen that the Pittsburgh and AVEC2014 datasets are not very challenging due to the simple approaches used for face detection, such as Viola & Jones, Dlib and AAM in AVEC2014 and AAM in Pittsburgh. These approaches cannot detect faces in or near profile. In addition, AAM is not a practical approach because it is participant-dependent. That is to say, this approach needs to create a model and train it for each participant, which may require manual editing of the landmarks in the training process. Thus, this approach is not fully-automatic. The use of two approaches (Dlib and Openface) for face detection on in-the-wild data has been investigated in this thesis, described in Chapter 4. The findings showed that using Dlib landmarks gave better performance than Openface landmarks. A possible explanation for this might be that Dlib only detects the face when it is frontal or semi-frontal otherwise, it loses track of the face. However, Openface detects all the frames. This led to conclude that using the frames where the face is frontal could achieve better results than using all the frames, including the noisy ones. According to these results, the performance of a system may vary according to the level of variability in the dataset. Therefore, it is possible that these results may not be reproducible on a dataset recorded in the wild.

Table 3.7 Summary of video data information for participants with health conditions. A check-mark is used as an indication of whether the dataset is in the wild based on the authors' claim (W: In-the-wild, FPS: frame per second).

Data	Population (control)	Ground truth	Resolution (FPS)	Availability	W
BlackDog	60 (30)	Clinical assessment	800x600 (24.94)	Private	-
ORI	8 (4)	-	- (30)	Private	-
AVEC2013	292 (-)	Self-report	640 x 480 (30)	Public	-
AVEC2014	84 (-)	Self-report	640 x 480 (30)	Public	-
DAIC-WoZ	189 (-)	Self-report	-	Public	-
CHI-MEI	26 (13)	Clinical assessment	640x480 (30)	Private	-
Pittsburgh	49 (-)	Clinical assessment	640x480 (29.97)	Public	-
BD	95 (46)	Clinical assessment	- (30)	Public	-
OU2016	18 (9)	Clinical assessment	-	Private	-
OU2017	29 (15)	Clinical assessment	- (30)	Private	-
OU2019	24 (12)	Clinical assessment	- (25)	Private	-
IVA_{18}	18 (0)	Clinical assessment	Variable	Private	✓
IVA_{52}	52 (23)	Clinical assessment	Variable	Private	✓

Table 3.8 shows a review of only these two datasets because other datasets are private for ethical reasons and hence not available for other researchers to use for evaluation, as shown in Table 3.7. In addition, DAIC-WOZ provides only feature sets for a researcher for public use due to ethical reasons preventing them from sharing the video data.

3.3 Discussion and Limitations

An initial objective of the research in this thesis was to identify the meaning of the term 'in-the-wild data' by investigating what kinds of challenges have been reported by researchers in relation to commonly used datasets for healthy individuals and datasets for people with health conditions, as shown in Table 3.15 Section 3.5.5. Some of these datasets are referred to as 'in-the-wild', and others, even if authors have not labelled them as 'in-the-wild', are still challenging relating to the nature of the datasets' recording environment and the challenges included. This section will discuss the datasets described above according to the data type.

Table 3.8 Comparison between different face detector approaches and the performance of the depression detection obtained grouped by datasets (AAM: active appearance model, KLT:kanade-tomasi-Lucas Tracker).

Data	Paper	Population	Face Detector	Performance in Accuracy (or otherwise noted)
Pittsburgh	(Alghowinem et al., 2015) (Dibeklioğlu et al., 2015)	38(19) 48 (-)	AAM ZFace	recall=94.7% 86.21%
	(Dibeklioğlu et al., 2017)	48 (-)[3]	ZFace	72.59%
	(Cohn et al., 2009)	107(41)	AAM	76%
	(Alghowinem et al., 2020)	38 (19)	AAM	86.8%
AVEC2014	(Valstar et al., 2014)	50 (25)	Viola & Jones	82%
	(Senoussaoui et al., 2014)	50 (25)	Viola & Jones	82%
	(Alghowinem et al., 2015)	32 (16)[4]	AAM	recall=68%
	(Pampouchidou et al., 2016)	200 (166)[5]	KLT	74.5%
	(Alghowinem et al., 2020)	32 (16)	AAM	87.5%
	(Jan et al., 2017)	300	DL	MAE=6.68
	(Zhou et al., 2019)	300	Dlib	MAE=6.21

3.3.1 Commonly Used Data

For the more commonly used dataset, three different types of tasks are mostly considered: face detection and tracking, eye blink detection and head movement estimation. Below, the limitations for each of these types of dataset tasks are discussed.

The face detection datasets are mainly used for training models, not for testing systems. The face poses variations are mostly frontal, with few images in which the face pose reaches a maximum of $+30^\circ$ or a minimum of -30° degrees. It is interesting to note that, in all these datasets, one or multiple filtering approaches, which are automatic, manual or both, are used to detect the face. This filtering approaches use algorithms (e.g., Viola & Jones) which do not detect the face when it is in or near profile (Belhumeur et al., 2013; Huang et al., 2008) or when it is not greater than 500 pixels in width (Le et al., 2012). The most obvious finding to emerge from the analysis is that these datasets are used to detect facial landmarks and not track them within video frames. These datasets are available for public use.

Face tracking datasets have been utilised for face tracking and recognition tasks, not facial landmarks detection. Thus, the annotation of the facial landmarks is only available in some of them. According to the authors' claims, most of the datasets mentioned above are considered in-the-wild data, except for the RU-FACS data, which is reported as having been recorded in a lab-controlled environment, the NDS and the DDF datasets, which are not categorised to be in-the-wild. The video length of the datasets is less than one minute, which is very short. Previous work has not provided more details regarding the participants, environment and device used, which would have helped to show the reality of the challenges included in these datasets to allow for a fair comparison. In addition, the videos used for performing facial landmark detection and tracking are very short, between 2 and 3 seconds for most datasets and 1 minute for 300-VW data. This is because annotating facial landmarks in videos is high-cost in terms of time and effort (Sagonas et al., 2013a).

In terms of the eye blink detection datasets, the previously mentioned datasets varied in terms of the device used to record the data but not significantly, as the typical setup involved a single camera monitoring the participant's face. Notable exceptions were the setup employed for ZJU, Eyeblink8, and Basler5, which used an external camera to capture high-quality videos. Regarding the participant behaviour during the recording, the participants mostly were frontal the entire time. Table 3.4 presents every study that has assessed its eye blink detection approach grouped by the dataset used. It can be seen that most of the previous work used the Viola-Jones algorithm (Viola and Jones, 2004) to detect and track facial landmarks. This gives an indication of the simplicity of predicting the facial landmarks from these datasets. In terms of dataset availability, the Talking face dataset is the only one fully available for free download, whilst the rest of them need a signed End User License Agreement (EULA) in order to get access for download.

Finally, datasets for head movement estimation are considered more challenging than previous types of datasets regarding face pose variations. Most of these datasets have used external devices from a commercial sensor, such as the Kinect with a magnetic tracker transmitter attached to the participant's head (BU, ICT) or a template-based head tracker (BIWI). The specifications of the challenges for most of them are not clear. For example, the ICT-3DHP dataset only mentioned the camera used to capture the participants and how the data is annotated. In addition, the BU dataset lacks facial occlusion and contains no information regarding the participant's behaviour and the recording environment because the focus was only on the time-varying illumination challenge. Although the AFLW dataset is the only one that describes different challenges in their datasets, it is collected from the Internet and not recorded as a video, like the previous datasets. One of the main limitations of that dataset is that there is no annotation for the image if a facial landmark is not visible to the annotator. Taking all the above together, it is difficult to compare these datasets due to the lack of information about the recording environment and the participants' behaviour during the recording.

3.3.2 Health Conditions Data

The previous section discussed the limitations of collecting healthy individuals' data for evaluation based on the purpose of the dataset. Such datasets are mostly collected from the internet or in a lab-controlled environment. Therefore, those datasets were discussed individually according to their purpose. On the other hand, data collection of participants with health conditions, such as CI and depression, is a very challenging procedure of health conditions detection research because it often requires recruiting participants from hospitals and psychological or memory clinics. This section will discuss the datasets shown in Table 3.7 based on the recording equipment used, the sample size, their availability and the country, which were collected as follows:

Recording Equipment: The type of recording equipment varied due to the different organisations and labs that recorded these datasets. The difference is insignificant because the set-up usually involved one camera capturing the participant's face or body. However, only the Pittsburgh dataset showed an exception by employing four hardware-synchronised analogue cameras, two for capturing the head and shoulder of the participant, one for full body recording and a fourth for monitoring the interviewer activity with two microphones for speech recording.

Sample size: One of the important aspects of a dataset is the number of participants. Most of the datasets for people with health conditions are relatively small, which is a common problem and explained by the fact that collecting data for participants with memory problems and dementia is very challenging. Only the AVEC2013 consists of 292 participants, but its assessment was based on self-report, not clinical assessment.

Availability: In terms of the dataset availability, only the AVEC2013 and AVEC 2014 are open for researchers to use in depression recognition studies. The DAIC-WoZ and Pittsburgh are partly available for public use. All these datasets require a signed EULA to gain access for download. However, most of the health conditions datasets have not been released publicly due to privacy issues in the ethical guidelines.

Worldwide: Most of the datasets for people with health conditions were collected in the USA, Europe and Japan. These datasets involved one or more modalities (audio, video, and text). A considerable amount of literature has been published on depression datasets using different modalities and the combination of modalities, which is audio and video, due to the valuable information that can be obtained from videos (He et al., 2022; Pampouchidou et al., 2017). However, in the literature on dementia detection, most of the work has been conducted using two modalities, audio and text. Only one Japanese research group has investigated the use of video modality. To the best of our knowledge, this thesis is the first to explore the video modality for CI detection in English.

Taking together all the factors mentioned above, most studies in the field of computer vision for healthcare applications have focused only on evaluating their approaches on a dataset recorded in a lab-controlled environment. Therefore, their approaches may not achieve good performance if applied to the in-the-wild dataset, and as a result, the incapability of their approaches to be generalised. The effect of these approaches on data recorded in the wild remains unexplored. This current research aims to determine the extent to which it is feasible to extract facial cues from data recorded in the wild for people with CI and whether this can help to develop a home-based application in the future.

3.4 In-the-wild Data

A considerable amount of the relevant literature described in Chapter 2 has used data recorded in a lab-controlled environment to evaluate work. Acquisition of such datasets in particular environment settings has advantages for certain research areas, such as enabling the researchers to control the variability of challenges in the dataset. In addition, researchers in machine learning focus on developing their models to improve the performance of such datasets. However, one of the main challenges for deploying models or approaches in different settings, such as agriculture or healthcare, from the lab environment to the real world is the lack of data with high and representative variability.

Sagar (2021) suggested that the focus for improving the performance of their models should be data-centric rather than model-centric because data quality determines a good model or approach. To study more general and unconstrained approaches in terms of the computer vision field, videos or images of faces have to be gathered from highly diverse sets. Although gathering data with a large number of variables in the lab in an attempt to build such a dataset is possible, there are two main barriers to this approach. The first is that the procedure for collecting such data is tremendously intensive. The second is the difficulty of establishing which distributions of various parameters should be used to create the most useful data.

The term ‘in-the-wild data’ has been used in many studies, particularly facial landmarks detection and its related features, due to the need for a face detection model that can capture faces in different cases. Huang et al. (2008) included images of people with variations in age and a small variation in gender and ethnicity, a few images with poor lighting conditions, and small occlusions. Le et al. (2012) collected images with noisy backgrounds, facial expressions, and variations in clothing and ethnicity. Belhumeur et al. (2013) gathered data with images of variability, such as lighting, resolution, small occlusion, age, hairstyle and ethnicity. Sagonas et al. (2013a) collected 300 indoor images and 300 outdoor images with keywords like ‘party’, ‘conference’, ‘protests’, ‘football’ and ‘celebrities’. Koestinger et al. (2011) gathered data that exhibited variations in age, ethnicity, gender, clothing, hairstyles, facial expressions, and near-frontal and non-frontal faces. These experimental datasets are rather controversial due to the lack of agreement on the kind of challenges to include in in-the-wild data.

Research on this subject has been primarily restricted to limited challenges mostly related to environmental conditions and people’s demographics (e.g., age and ethnicity). There remains a lack of variety in other aspects related to the participants’ behaviour and environmental conditions. This chapter suggests that ‘in-the-wild data’ should cover many of the edge cases of the participants in terms of demographics, look and behaviour, environment conditions and consumer device used. The challenges regarding these participants, environment and devices are summarised in Table 3.9.

The participant-based challenges are divided into demographics, look, and behaviour. Demographic includes variations in age, gender and ethnicity. The variation in age consists of children, young and young adults, adults and older people, depending on the purpose of the dataset. Participant look consists of variations in clothing, hairstyle, makeup style and glasses style. The glasses’ style comprises glasses reflection, frames and sunglasses because reflection and frames may cause a problem in facial landmarks detection, especially those in the eye region. For example, laptop and light reflections can appear on glasses, and if the glasses’ frame is situated in the middle of the eyes, it makes it difficult to capture the eye activities. In addition, sight sunglasses may affect performance in the prediction of facial landmarks in the eye region due to the dark colour of this kind of glasses. Participant behaviour includes facial expressions, the participants sitting in a non-optimal position with respect to the camera and the light, a varied distance to the camera over time, spontaneous small and large head movements, spontaneous body movements and hand movements on the face. This behaviour is not commonly captured because researchers mostly depend on a lab-controlled environment for video recording.

Most challenging datasets are built based on the environment-based aspect, which is the most common one. This aspect includes low resolution, poor illumination, indoor or outdoor environment and a noisy background in which furniture, animated cartoons, pictures and people can appear behind the participant. A further challenge is the appearance of more than one person in front of the camera if these people are closer to the camera than the

Table 3.9 Potential challenges in in-the-wild dataset.

Participant-based
Demographics
(1) Variation in age
(2) Variation in gender
(3) Variation in ethnicity
Look
(4) Variation in clothing
(5) Variation in hairstyle
(6) Variation in glasses style
(7) Variation in makeup style
Behaviour
(8) Facial expressions
(9) Non-optimal position with respect to the camera
(10) Varied distance to the camera over time
(11) Non-optimal position with respect to the light
(12) Spontaneous small head movement
(13) Spontaneous large head movement
(14) Spontaneous body movement
(15) Hand movement on the face
Environment-based
(16) Low resolution
(17) Poor illumination
(18) Indoors environment
(19) Outdoors environment
(20) Noisy background (e.g., devices, animated carton, pictures, people and furniture)
(21) More than one person in front of the camera
Consumer Devices-based
(22) Smartphone camera
(23) Laptop camera
(24) Professional camera

participant. The final aspect is device-based, which consists of smartphones, laptops and professional cameras. Most related studies have used professional camera devices to record videos. However, smartphone and laptop webcams can result in considerable challenges of the type mentioned above.

In this section, the aspects of ‘in-the-wild data’ have been explained. The following section will describe the datasets used in this thesis and then use the summary of the challenges reviewed of in-the-wild data to examine this thesis’s datasets and a range of previous studies’ datasets.

3.5 Data Used in this Research

This section first describes what type of datasets are used in this research and then why these datasets can be considered in the wild.

3.5.1 Task

This study uses data provided by the Hallamshire Hospital Memory Clinic in Sheffield, UK. Details of the experiment were given to every participant when he/she agreed to participate. The videos are recorded using a laptop camera or a smartphone to capture each participant’s face (and/or the accompanying person) (see Figure 3.6). The audio is recorded using $Tascam^{TM}DR-40$, which was placed on a table, and two microphones were attached to the participants. The recorded videos are for people with different types of CI, MCI and neurodegenerative disorder (ND), functional memory disorder (FMD) and healthy controls (HC), as they answer memory-probing questions asked by an intelligent virtual agent (IVA). The questions are of different types: open questions, closed questions, and compound questions to asses participants’ long and short-term memory. The diagnostic details for every participant are provided. Ethical approval for collecting and using this data was given by the National Research Ethics Service (NRES) Committee South West-Central Bristol (Rec number 16/LO/0737) in May 2016.

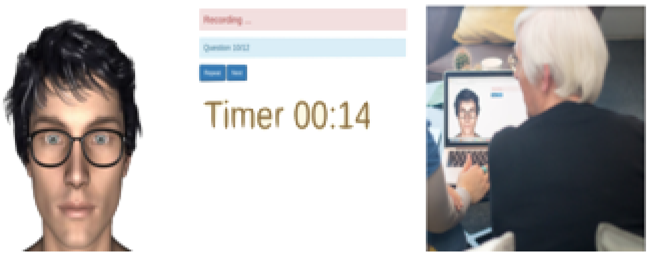


Fig. 3.6 A screen-shot that presents the IVA when it is in use (Mirheidari et al., 2018).

Table 3.10 shows the number of participants recorded every year for each group (i.e., ND, MCI, FMD and HC). Due to the difficulty of recording many participants, this research initially used 18 participants from IVA2016 because this was the only data provided by the memory clinic at that time. This data will be referred to as IVA_{18} throughout the thesis. Then, 34 more participants were provided in 2021, including recordings from 2017 to 2021. This data is referred to as IVA_{34}. This data includes 23 HC but very few participants with health conditions (i.e., 5 with ND, 4 with MCI and 2 with FMD), making it difficult to use the data independently for the evaluation. Therefore, the IVA_{34} data is combined with the IVA_{18} data, which results in a larger dataset with 52 participants, which is referred to as IVA_{52}. The following sections will describe these three datasets individually.

Table 3.10 The number of participants in different versions of the IVA (2016/2017/2018/2020/2021) datasets with their diagnostic classes. (ND: neurodegenerative disorder, MCI: mild cognitive impairment, FMD: functional memory disorder and HC: healthy controls).

	ND	**MCI**	**FMD**	**HC**	**Total**
IVA2016	6	6	6	0	18
IVA2017	0	0	2	0	2
IVA2018	3	3	0	2	8
IVA2019	2	0	0	1	3
IVA2020	0	0	0	8	8
IVA2021	0	1	0	12	13
Total	11	10	8	23	52

3.5.2 IVA_{18} Dataset

This data includes a total of 18 participants who were recorded in 2016 (IVA2016), split equally into 6 with ND, 6 with MCI and 6 with FMD. 4 were excluded because they have depressive pseudodementia and 2 for whom the diagnosis was not clear. All participants are in the age range 43 to 78– for more information about the demographic information, see Table 3.11. The duration of the videos in total is 208 minutes (mean = 11.56 minutes). The dataset is small, but this is a common issue in similar studies that involve human participants in clinical settings, as explained previously. The participants were told that they could bring someone with them, and, as a result, 6 of the 18 participants brought an accompanying person with them (4 ND, 1 MCI and 1 FMD) (see Table 3.12). Therefore, some videos contain four people: the participant, the accompanying person, the neurologist, and the person who operates the laptop. Although the participants were not given any specific instructions as to where to look, the talking head on the screen will have been the most salient point on which to look. This poses a challenge for video-based processing. The study under which the recordings were done mostly focuses on speech processing and is the first research study to be done using videos. It is worth mentioning that previous studies have focused more on distinguishing dementia (regardless of the dementia type) from the HC group using video data. On the other hand, this research differentiates between the three types of memory problems (ND vs MCI vs FMD) as discussed in Chapters 4 and 6.

Table 3.11 Demographic information of the IVA2016 participants.

	ND (n=6)	**MCI (n=6)**	**FMD (n=6)**	**Total (n=18)**
Age	65.8 (+/-10.38)	63.3 (+/-8.96)	55.7 (+/- 8.94)	61.6 (+/-10.16)
Female	33.3%	33.3%	16.7%	27.8%

Data Challenges

When this data was recorded, it was not intended for video processing purposes. That is why this data contains a high level of noise due to the lack of restrictions on the participants and the environment with respect to the webcam position. In this research, this data is referred to as in-the-wild data because it includes more conditions related to in-the-wild data, such as a semi-dark or dark and noisy room, as will be discussed in Section 3.5.5. In addition, spontaneous behaviour means that participants can act as they would in their natural environment, such as moving about freely. A participant can continually change the orientation of their face, rotate their body, and move closer to and further away from the camera. Other people can also appear with the participant and move around too. In addition, participants who wear glasses sometimes have their eyes obscured by the frames or a reflection from the laptop on the glasses. The majority of the IVA_{18} data recordings were recorded at 30fps. However, five recordings were done at 24fps, producing a different resolution recording. These issues cause complications for automatic methods to extract visual information from the data.

3.5.3 IVA_{34} Dataset

Data is collected in two different environments, a clinic and at home. People at home used laptops and smartphones to do the recording. There are 34 videos (19 female and 15 male) representing four groups: 5 participants with ND, 4 with MCI, 2 with FMD and 23 with HC. The total duration of the videos is 538.59 minutes (mean = 9.79 minutes, SD = 5.54

Table 3.12 Details about the participants of the IVA_{18} data by showing whether the participant came with an accompanying person and on which side this person sat (on the right- or left-hand side of the participant).

Participant	Gender	Age	Condition	Accompany?	Which Side?
P01	M	43	FMD	No	-
P06	M	57	FMD	No	-
P07	F	45	FMD	Yes	Right
P08	M	65	MCI	No	-
P09	M	67	FMD	No	-
P10	M	58	FMD	No	-
P11	F	52	ND	Yes	Right
P12	M	63	MCI	No	-
P13	M	69	ND	No	-
P14	M	51	MCI	NO	-
P15	F	78	MCI	Yes	Right
P16	F	78	ND	Yes	Right
P17	M	64	FMD	No	-
P18	M	58	MCI	No	-
P19	F	65	MCI	Yes	Right
P21	M	76	ND	No	-
P22	M	63	ND	Yes	Left
P23	M	57	ND	Yes	Left

minutes). Eight participants were excluded from the study due to the difficulty of face detection resulting from several challenges, such as a dark room and the partner interrupting the participant multiple times during the session, the eyes of the participants not being visible to the camera, two of them were wearing masks and the non-optimal angle of the participant from the smartphone camera. Moreover, participants looked and talked most of the time to the right side at the person who was operating the laptop.

As with the IVA_{18} data, in the clinic recordings, participants were told that they could bring someone with them. Four out of the 13 participants recorded in a clinic brought a caregiver or partner with them (3 ND and 1 MCI), in which all of them were female.

Some videos, consequently, contain four people, as mentioned in the previous section. In the home recordings, only 2 out of the 20 participants who made a home recording had a caregiver/partner with them during the session (1 MCI and 1 HC). As the participants were in their homes, the setting environment for the recording was either an office, a living room, or a bedroom. Every participant shows a different challenge for the face detector because he/she had the choice of which room to sit in, the distance and the angle from the camera, the lights of the room on or off and whether to do the session during the day or at night time.

Tables 3.13 and 3.14 show the participants' information, such as who was accompanied, where his/her accompanying person sat, condition, age and gender. Of the seven participants who were accompanied, four had ND, two had MCI, and one was from HC. According to Larner (2012), Larner (2014b) and Soysal et al. (2017), coming with a partner or other person and head-turns might be clinical cues of having a CI. This also has been shown in experiments described in Chapter 6.

Table 3.13 Details about the participants of the IVA_{34} data that are recorded in the clinic.

Participant	**Gender**	**Age**	**Condition**	**Accompany?**	**Which Side?**
P39	M	50	FMD	No	-
P40	F	58	FMD	No	-
P57	M	72	ND	Yes	-
P82	F	56	MCI	No	-
P84	F	67	MCI	Yes	Left
P89	M	55	HC	No	-
P98	F	63	ND	Yes	Left
P115	F	64	ND	Yes	Right
P116	M	Unknown	HC	No	-
P117	M	Unknown	ND	No	-
Patient 1	F	Unknown	MCI	No	-
Patient 2	M	Unknown	HC	No	-
Patient 4	F	78	ND	Yes	Right

Table 3.14 Details about the participants of the IVA_{34} that are recorded at home.

Participant	Gender	Age	Condition	Accompany?	Which Side?
P219	F	Unknown	HC	No	-
P222	F	Unknown	HC	No	-
P223	M	Unknown	HC	No	-
P226	F	Unknown	HC	No	-
P230	F	Unknown	HC	No	-
P261	M	68	MCI	Yes	Right
P264	F	Unknown	HC	No	-
P265	M	Unknown	HC	No	-
P268	F	Unknown	HC	No	-
P269	M	Unknown	HC	No	-
P270	F	Unknown	HC	No	-
P274	F	Unknown	HC	No	-
P276	F	Unknown	HC	No	-
P282	F	Unknown	HC	No	-
P350	F	Unknown	HC	No	-
P355	M	Unknown	HC	No	-
P356	M	Unknown	HC	Yes	Right
P357	M	Unknown	HC	No	-
P358	F	Unknown	HC	No	-
P359	M	Unknown	HC	No	-

Comparing IVA_{34} Data Challenges to IVA_{18} Data

As previously stated, the IVA_{18} data contains many challenges, such as low resolution, different illumination, the appearance of many people in front of the camera, sitting at a different distance from the camera, body movements and large head movements. These challenges could also be seen in the extended data. The IVA_{34} data is more challenging than IVA_{18} because the recordings were made at home and people used different devices (laptops and smartphones). That is to say, the participants had more freedom and were more comfortable choosing any room, sitting in a room while the lights were turned off, recording the session while the day/room lights were behind the participant, and noisy backgrounds, including furniture such as the office table, library, pictures of people, and antiques. Setting the laptop at any angle without any guidelines made one of the participants sit at a wrong angle, make a face not facing the camera, or make part of the face not visible to the camera. Some home-recorded videos were recorded using a smartphone, meaning the participant may have a different angle view to the camera. In laptop recording, people usually sit in front of the laptop or at an angle from the laptop. However, the eye's position is looking toward the camera. On the other hand, people who use their smartphones hold their phones at a lower angle from their face, which makes their eyes look partly closed, and their head orientation is different. In addition, people at this age need to wear eyeglasses, which also causes problems. These challenges may lead to many problems that are resolved by removing some videos, as explained above, and cropping some of them to exclude other people who appear closer to the camera than the participants.

Although clinic recordings and home recordings share several challenges, the challenge of people appearing with the participant in the camera view is more common in the clinic recording than in home recordings. In the clinic recordings, at least two people appear on camera: the person who operates the laptop and the participant. Two additional people may appear on camera, as previously mentioned, the doctor and the partner, resulting in four people visible in the video.

3.5.4 The IVA_{52} Dataset

Due to the small number of participants with health conditions in the IVA_{34} data, the IVA_{18} and IVA_{34} are combined to form the IVA_{52} dataset to evaluate the proposed approaches in a larger dataset with a HC group. Table 3.10 shows the number of participants from the different IVA datasets versions for each class and the total number of participants, which is 52, split into 11 with ND (45.5% female), 10 with MCI (50% female), 8 with FMD (25% female), and 23 HC (52% female). Since the IVA (IVA2017/18/19/20/21) includes questions that were not in the original version of the IVA2016, only the same asked question amongst all the versions are included in this study.

3.5.5 Comparing this Research's Dataset Challenges with Previous Work's Dataset Challenges

Table 3.15 shows the previously reviewed datasets organised according to their purpose in rows and the in-the-wild data aspects based on participant, environment and device used in columns. The reported challenges of these datasets are indicated by a check-mark. The check-mark is an indicator of the presence of this challenge in the dataset, not an indicator of the degree of the challenge. For example, when a particular dataset includes more than one particular ethnicity, it is check-marked as the same as a dataset with several ethicities. In Table 3.15 below, the ‘W’ column contains a check-mark if that particular dataset has been claimed to be in the wild by its authors. The main points of this table are summarised as follows:

- Of the different aspects of in-the-wild data, most researchers have focused on the recording environment by including images or videos with different illumination and noisy background. However, a few of them have low image quality. The researchers have collected challenging images, but Sagonas et al. (2016) suggested using videos instead of images, which are more complicated to handle due to the unexpected behaviour of the participants.

- In addition, most of the datasets are recorded in an indoor environment. However, only some of them have recordings from both indoors and outdoors. These studies claim to be in-the-wild data because of include images or video clips from indoor and outdoor environments, which indicates the significance of this factor in creating this kind of dataset.

- In the participant-based aspect, previous work has mainly focused on three challenges: gender from the demographic category and facial expressions and spontaneous small head movement from the behaviour category. Most of the previous work that included facial expressions as one of the challenges have only "smile" expressions.

- Moreover, the small head movements are mostly not spontaneous, as people were asked to move their heads for the video recordings, and the rest were images, and as mentioned previously, videos are more challenging than images. Moreover, few studies have focused on collecting data from people wearing glasses.

- Interestingly, no previous work has collected data for people with different behaviour, as presented in the table. Previous datasets consisted of noisy backgrounds, such as pictures, furniture or people, but no data had the issue of having other people closer to the camera than the participant. Although a few datasets had people wearing glasses, glasses-related issues have not been explored.

- In the device-based aspect, only one study mentioned that they used a laptop to record the sessions. In contrast, other datasets have been collected from the internet as images or using an external camera for the videos.

- It is clear that these datasets are limited due to the lack of variations in the population's age and ethnicity.

- According to the 'W' column, nine datasets are claimed to be considered in the wild. Figure 3.7 shows an analysis of the common challenges in these datasets based on their authors regardless of the data type (i.e., images or videos) in the upper figure, which resulted in 15 challenges. It can be seen that most of these datasets' authors focused mainly on including particular challenges during the data collection procedure, which are variations in ethnicity, illumination, gender, clothing, hairstyle, facial expressions, indoors and outdoors, and small head movement.

- Then, another analysis is conducted in order to compare these datasets in terms of image-based data (7 datasets) and video-based data (2 datasets), as shown in the bottom figure of Figure 3.7. According to the figure, image-based datasets included more challenges than video-based datasets. The authors of the video-based datasets are primarily interested in collecting datasets with variations in age and illumination, wearing glasses and sunglasses, and small and large head movements.

- The more reasonable explanations of why video datasets lack many challenges are 1) the limited number of these datasets, 2) the videos are short, lasting from 3 to 60 seconds, and 3) the difficulty of annotating video data frame-by-frame.

- The IVA data used in this research covers most of the challenges described in Section 3.4. This data can be considered more challenging than previous in-the-wild video-based datasets because the videos are very long compared with the previous work's datasets and are recorded without any restriction on the participants or the environment. For these reasons, participants can exhibit more spontaneous behaviour during the session. The recordings are not made in a lab-controlled environment.

- In addition, Asgarian et al. (2019) and Taati et al. (2019) revealed that the performance of state-of-the-art facial landmarks methods can be affected (less accurately or fail) when it is evaluated on people with CI rather than healthy people. This provides an insight into the limitation of these methods on the clinical population.

- Finally, it is very hard to build this kind of in-the-wild dataset due to the previously mentioned points and procedures required to do such a task, especially recording videos, because that comes with a high cost in time and effort.

Table 3.15 Presenting the challenges of every data commonly used for evaluation purposes using healthy individuals or people with health conditions (W: In-the-wild, FD: Face Detection, EBD: eye blink Detection and HME: Head Movement Estimation).

Purpose	Data	W	Participant-based															Environment-based						Device-based		
			Demographic			Look				Behaviour																
			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
FD	Helen	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓		✓				✓	✓	✓				
	LFPW	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓			✓	✓	✓	✓				
	LFW	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓		✓		✓		✓	✓	✓				
	300-W	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓		✓				✓	✓	✓				
	IJB-FL	✓		✓	✓	✓	✓	✓		✓				✓	✓			✓	✓	✓	✓	✓				
	Menpo	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓			✓	✓	✓	✓	✓		
	300-VW	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓				✓
	YouTube Celebrities	✓	✓	✓	✓		✓	✓			✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓			
	DDF													✓	✓				✓	✓	✓					
	NDS							✓						✓	✓			✓	✓		✓					
	RU-FACS																									
EBD	ZJU							✓												✓		✓				
	Talking face									✓										✓						
	EyeBlink8			✓						✓										✓						
	Basler5																			✓						
	RN					✓	✓	✓						✓			✓		✓	✓		✓				
HME	AFLW	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓			
	BU													✓					✓							
	ICT-3DHP													✓												
	BIWI			✓									✓				✓		✓							
Depression	BlackDog			✓						✓				✓						✓						
	ORI			✓																						
	AVEC2014		✓	✓								✓		✓				✓	✓	✓		✓			✓	
	DAIC-WoZ																									
	CHI-MEI																									
	Pittsburgh		✓	✓																						
	BD		✓	✓																						
Dementia	OU			✓						✓										✓						
	IVA$_{18}$	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓		
	IVA$_{34}$	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	

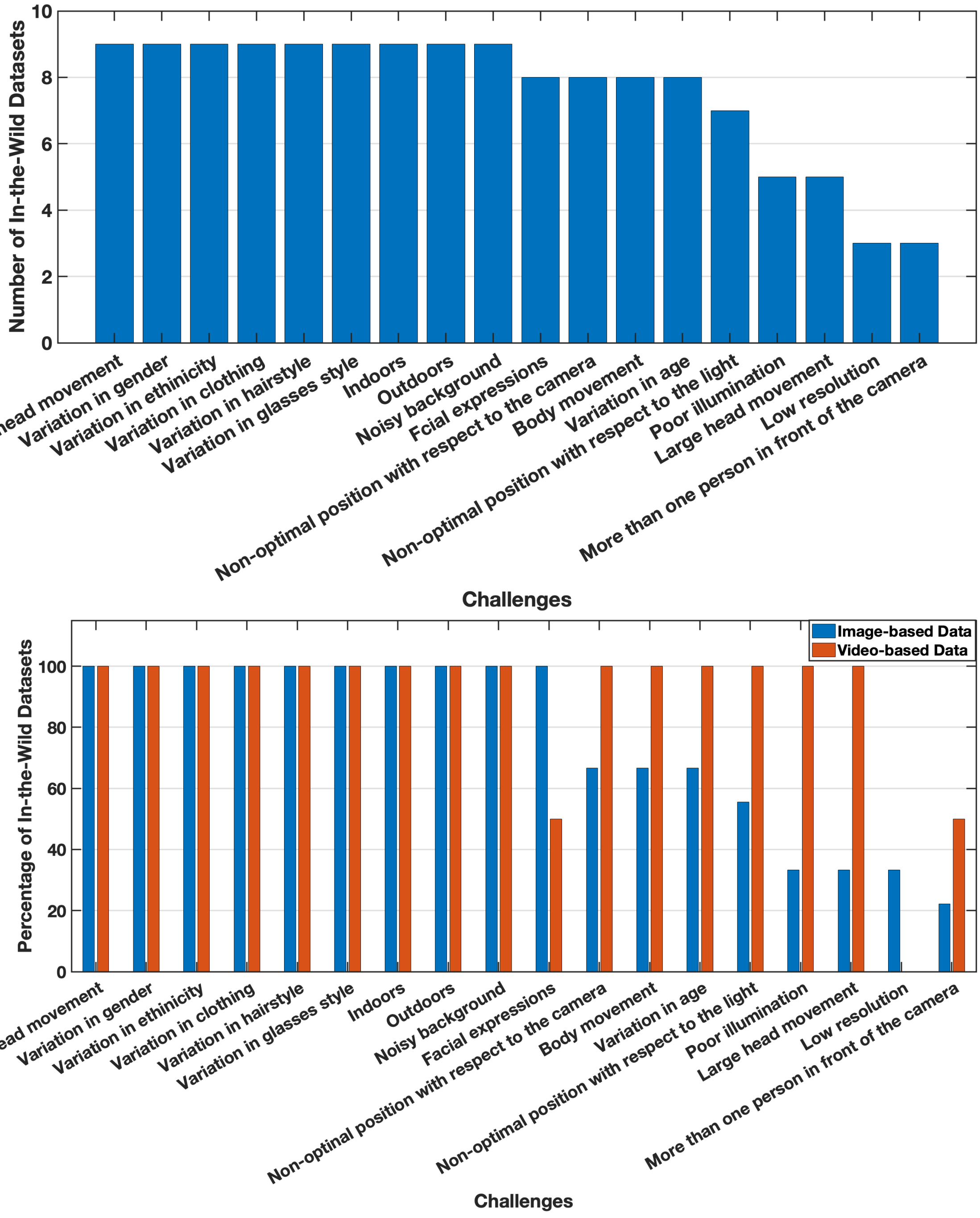


Fig. 3.7 Analysis of the common challenges of all the in-the-wild data according to these datasets' authors is shown in the upper figure, and the challenges that are image-based or videos-based are shown individually in the bottom figure.

3.6 Summary

The purpose of this chapter was to determine what kind of conditions should be included in the data for it to be considered as in the wild. This chapter then reviewed different dataset types: commonly used datasets for both healthy individuals and people with health conditions, including their limitations. Then, the datasets used in this research were introduced, including their challenges. A comparison of the challenges of this research dataset with the challenges in previous work datasets was then made.

This work is the first to investigate this problem and lay the groundwork for future research into collecting datasets based on in-the-wild conditions. This work has shown that in-the-wild data involves several aspects based on the participant's demographics, look and behaviour, the environment and the device used. The existence of these aspects may have variations in particular challenges. In addition, analysis of previous work datasets showed that each dataset has weaknesses and strengths based on the purpose for which it is used. However, this research has identified significant points which are similar for all the datasets regardless of the dataset type: 1) some datasets have been used for a different reason from the purpose for which they have been built, 2) there is a lack of information about the datasets in terms of participant, environment and device used and 3) most of them used professional devices to record the dataset.

In addition, this chapter has highlighted the most common challenges included in the datasets which were claimed by their authors to be in the wild and found a difference between the included challenges in terms of the type of the dataset (i.e., image-based or video-based).

Most researchers in the field of computer vision for healthcare applications have only focused on evaluating their works on data recorded in a lab-controlled environment. This work is the first to explore an in-the-wild dataset to detect particular health conditions. The dataset used in this research may provide insights into 1) the importance of investigating the detection of health conditions using datasets recorded in real-world scenarios and 2) developing approaches that have reliable performance on such data. These insights gained from this study could be of assistance to developing a home-based application in the future.

The rest of this research will be dedicated to investigating visual feature extraction, applying them on the IVA_{18} dataset, and validating them on the AVEC2014 dataset to compare the performance obtained in this research with related work. Finally, the methodology in this research will be applied to the IVA_{52} dataset of people with CI and FMD. Table 3.16 presents where each dataset will be used in the following chapters.

Table 3.16 Showing which data is used for each chapter.

	ZJU	IVA_{18}	AVEC2014	IVA_{34}	IVA_{52}
Chapter 4	✓	✓			
Chapter 5				✓	✓
Chapter 6		✓		✓	✓
Chapter 7			✓		

Chapter 4

Eye Blink Based Detection of Cognitive Impairment

"It's ok not to be okay, but it's not okay to stay that way."
— Perry Noble

4.1 Introduction

The previous chapter provided an analysis of different datasets recorded in a lab-controlled environment and in the wild. The analysis showed different kinds of challenges and how these challenges have a severe impact on computer vision techniques in terms of the capability for detecting facial features. This chapter will investigate an automatic approach to detecting eye blink rate (EBR) using two datasets: the ZJU and the IVA_{18} datasets. The ZJU is a public dataset used to evaluate the baseline system and compare the performance obtained with related work. Then, the IVA_{18} is used to explore the feasibility of this approach in identifying CI and related health conditions using the IVA_{18} dataset.

Eye blink is a major area of interest within the fields of security (Chen and Amayeh, 2019; Seha et al., 2019), human-computer interaction (Grauman et al., 2001), and assistive technology and healthcare (Argilés et al., 2015; Bentivoglio et al., 1997; Chermahini and Hommel, 2012; De Jong and Merckelbach, 1990; De Padova et al., 2009; Sun et al., 1997; von Cramon and Schuri, 1980). A range of techniques have been developed for detecting eye blinks, such as sensors (Al-Rahayfeh and Faezipour, 2013b), neural networks (Anas et al., 2017; Cortacero et al., 2019; Fogelton and Benesova, 2018; Schillingmann and Nagai, 2015), and eye aspect ratio (Soukupová and Cech, 2016); see Section 2.4.1 for more details.

The process of eye blink detection is a very challenging task for machines, especially for in-the-wild data, because it can be adversely affected by conditions such as the recording environment, the devices used for recording, and the spontaneous behaviour of the participant. A more detailed account of these conditions is given in Section 3.4. Various studies have assessed the efficacy of eye blink detection using standard datasets, recorded in a lab-controlled environment, and have obtained high performance. Such approaches, however, may be unreliable precisely because they were conducted on data recorded in a lab-controlled environment. Much less attention has been paid to eye blink detection using data recorded in the wild.

Only a few studies have attempted to investigate the detection of cognitive impairment (CI) based on the visual modality (Tanaka et al., 2019, 2017, 2016). Moreover, no previous study which focused on detecting CI has investigated the EBR as a visual cue. As described in Section 2.2.2, the EBR is linked to CI. Ladas et al. (2014) reported that people with mild cognitive impairment (MCI) showed a higher rate of eye blinks than healthy people, and the increase in their EBR was found to be associated with an increased risk of transition from MCI to Alzheimer's disease. Since eye blinks are an important cue of CI, they could be affected by several factors related to ageing and environments (e.g., temperature, brightness, air conditions, and relative humidity) (Sun et al., 1997). This chapter proposes a new methodology for calculating the EBR for data recorded in the wild, which has not been examined previously.

Figure 4.1 shows the workflow for automatic CI detection. Prior to face detection and tracking, the pipeline includes the pre-processing phase for the video data to address some of the in-the-wild data challenges, which is explained in Section 4.2.1. Next, face detection and tracking are performed. Different features are then obtained from the video frames and given to a supervised machine learning classifier to distinguish between the health conditions. These classifiers are trained on a sample of the data and use the resulting model to classify the new sample set. Section 2.5 gives details about the classifiers used and the evaluation metrics.

The remainder of this chapter proceeds as follows. Section 4.2 describes the development of an automatic eye blink detection system. Section 4.3 covers the evaluation for the baseline and the proposed approach on the IVA_{18} dataset using a set of popular classifiers. Then, a performance analysis of individual classes is conducted in Section 4.3.3.

4.2 Approaches to the Automatic Detection of EBR

These experiments examine the feasibility of automatically extracting eye blinks to detect CI from IVA_{18} data. This aim involves developing a technique designed specifically for in-the-wild data. The first and foremost step in achieving this aim is to develop a baseline

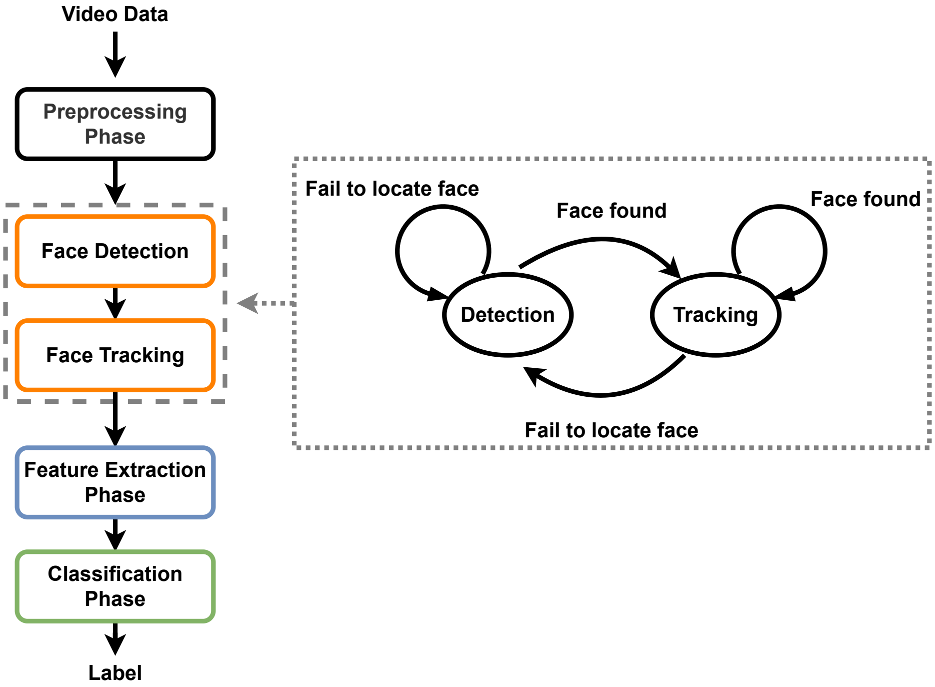


Fig. 4.1 An overview of the automatic CI detection workflow, which includes the state transitions for the face detection and tracking phase.

approach for detecting eye blinks and then evaluate the approach on standard data. The baseline results could give an insight into the difficulty level of eye blink detection and the classification problem with different classifiers using the same feature set.

As has been discussed in Chapter 2, although many researchers have utilised data recorded in a lab-controlled environment to measure the performance of their approaches to eye blink detection, their work may not perform as well on challenging data, like the IVA_{18} dataset, described in Chapter 3. This chapter proposes a novel approach to detecting eye blinks using multiple thresholds (MTs) to calculate the EBR. This approach addresses the complexities of the eye blink phenomenon in in-the-wild data. These experiments were conducted on the IVA_{18} dataset. The data details in terms of participants, diagnostic classes, and recording settings can be found in Chapter 3.

This section describes two approaches to detecting EBR in a video frame: automatic calculation of a threshold (the baseline system) and MTs approaches. In addition, this section explores how these approaches are performed using two different facial landmarks tracking techniques.

4.2.1 Dataset Preprocessing

Data preprocessing involves data preparation, cleaning, normalisation, transformation, and reduction (e.g., feature selection). This section briefly describes how raw data is prepared to be usable and valuable for further facial feature extraction techniques. Prior to the feature extraction phase, the IVA_{18} videos are preprocessed due to some people's appearance with the participant in front of the camera. These people are sometimes closer to the camera than the participant. The process is conducted by cropping the height and width of the video frames to detect only the participant while keeping the background noise. The cropping operation resolves only one challenge and does not remove any other challenges to ensure the data can still be thought of as being in the wild. Then, two different facial landmark tracking techniques are used to extract the facial landmarks, which will be explained in the following section.

4.2.2 2D Facial Landmarks Tracking

Facial landmark detection is challenging due to several factors: 1) participant-based factors (e.g., facial expressions, facial occlusion and head poses), 2) environment-based factors (e.g., variation in illumination, low resolution and noisy background) and 3) device-based factors (e.g., laptop and smartphone), as explained in details in Section 3.4. Many researchers have used Dlib (King, 2009) and OpenFace (Baltrusaitis et al., 2018) to automatically detect the locations of the facial key landmark points on images or videos (Ringeval et al., 2018, 2019; Zhou et al., 2018). Both techniques can identify 68 facial key landmark points, as shown in Figure 4.2. This section gives an overview of these two approaches by comparing them in terms of functionality, their advantages and disadvantages and in what cases each is the best approach.

Dlib

Dlib [1] is used for facial landmarks detecting and tracking. This detector is based on a histogram of oriented gradients image descriptors and the linear support vector machine (SVM), which, as proposed by (Dalal and Triggs, 2005), can be trained as an accurate detector of human faces. The Labelled Faces in the Wild dataset (Huang et al., 2008) has been used to train Dlib using 2,825 images.

This algorithm has several advantages, such as fast computation on the CPU, detecting frontal and slightly non-frontal faces and working with small occlusions. However, several

[1] http://dlib.net

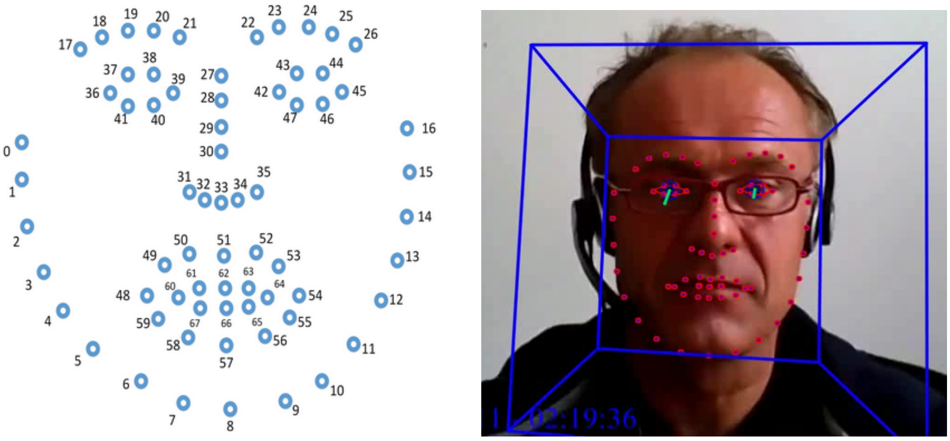


Fig. 4.2 The 68 facial landmarks defining the face shape (Baltrušaitis et al., 2016) (left) and a sample image with detected facial landmarks using Openface (right).

problems affect face detection using Dlib, such as poor illumination, low resolution, profiles of faces or extreme poses, small or far away faces and substantial occlusion (Gupta, 2018). Moving the laptop and screen angle variation can also cause problems. In addition, the dataset is limited to a few children, a few people over the age of 80, no babies, a few women and a few different ethnicities. This might affect the performance of the Dlib in face detection on datasets with high variations of people of different genders, ethnicities and ages.

OpenFace

OpenFace [2] is a toolkit for facial behaviour analysis. It detects and tracks based on the Convolutional Experts Constrained Local Model (Zadeh et al., 2017). In this chapter, it is used for face detection and tracking because it is robust at detecting facial landmarks in more realistic recordings with certain conditions, such as very poor lighting conditions and the inclusion of profiled faces.

For training the model, a number of datasets were used, such as LFPW, Helen, the Menpo, and Multi-PIE (Belhumeur et al., 2013; Gross et al., 2010; Le et al., 2012; Zafeiriou et al., 2017). Then, it was evaluated the model on two public datasets: IJB-FL (Kim et al., 2016) and 300VW (Shen et al., 2015). The IJB-FL (Kim et al., 2016) is a subset of IJB-A (Klare et al., 2015) and consists of 180 images (128 frontal and 52 profile faces). The 300-VW includes 64 videos. These datasets consist of images with faces in the wild (i.e. with variations in illumination, different face poses, even extreme ones, indoor and outdoor environments and different variations in the resolution of faces with slight and strong occlusions).

[2] https://github.com/TadasBaltrusaitis/OpenFace/

In this research, both techniques are used for the IVA_{18} dataset to evaluate the performance of the proposed approach when 1) only the frontal or semi-frontal facial images are used and 2) all the frames of the video, including the noisy ones and profile images, are used.

4.2.3 Feature Extraction

For over 40 years, many researchers have investigated approaches to eye detection and tracking (Hansen and Ji, 2009). However, this task remains challenging due to the fact that the eye appearance changes significantly across participants under different conditions, such as facial expressions, occlusion by an object or self-occlusion, low resolution and light conditions.

Neural network techniques have been employed to detect the eye status from images (Anas et al., 2017; Han et al., 2018; Li et al., 2018c). Anas et al. (2017) employed a convolutional neural networks (CNNs) approach to identify eye status (open/partially-opened/closed). They trained their model on the Helen dataset, which includes facial images of participants of different ages, genders, and ethnic origins (Le et al., 2012). In addition, the images have variable resolution, illumination, and pose conditions. However, the Helen dataset is very limited in closed-eye frames, which means that the eyes are mostly open. Additionally, they validated their work on the ZJU dataset. They used precision and recall metrics, which achieved 98% and 89.8%, respectively. Their work outperformed that of Kim et al. (2017), who also utilised CNNs to detect the eye status and used two combined datasets to evaluate their work: 1) their own dataset that was collected while the participant watched TV and 2) the ZJU dataset. They obtained an error rate of 0.23663% for detecting the eye status as closed or open. Furthermore, Han et al. (2018) also proposed a CNNs approach and evaluated their work on the ZJU dataset using precision and recall metrics which gave 94.4% and 89.7%, respectively.

A considerable amount of literature has been published on eye detection using different facial landmarks approaches (Baltrusaitis et al., 2018; Ouanan et al., 2016; Sagonas et al., 2013b). These studies showed the importance of these landmarks as an initial step for building or developing many applications ranging from biometric recognition to mental state understanding. Thus, various methods have been proposed for eye blink detection based on videos instead of processing an individual image to overcome the limitation of previous work (Appel et al., 2016; Lalonde et al., 2007; Soukupová and Cech, 2016).

Fogelton and Benesova (2016) proposed an algorithm to detect an eye blink using a state machine (SM) to determine the blink duration. For each eye, there is an SM, and a blink is considered if at least one of the SMs detects a blink. The intervals of the detected eye blink from both eyes are merged to improve the precision of the eye blink detection. This is done by calculating the intersection over union ($IOU = (A \wedge B)/(A \vee B)$, where A and B are the blink intervals for both eyes). An IOU value greater than 0.2 is considered a blink. When they evaluated their work on the ZJU dataset, they used the precision and recall metrics, which gave 99.2% and 97.3%, respectively. Then, they enhanced their approach by using dense optical flow to extract the feature and feed it to a recurrent neural network (Fogelton and Benesova, 2018). They achieved 97.6% as an F-measure metric.

Soukupová and Cech (2016) developed an algorithm to localise the landmarks of the eye region and calculate the eye aspect ratio (EAR), which was then used with the SVM to find the eye status. They found that rapid small head movements could result in false values for eye blink estimation, which indicates that even neural networks face issues in detecting eye status with small head movements.

Many studies have employed the use of the EAR for calculating the EBR and achieved very good performance, as described in Section 2.4, (Dewi et al., 2022; Maior et al., 2020; Navastara et al., 2020; Utaminingrum et al., 2021). Following previous work, this thesis is, therefore, adopted the EAR algorithm to estimate the eye blink (Soukupová and Cech, 2016). To calculate the EBR, six eye landmarks (x,y), as shown in Figure 4.3, are used.

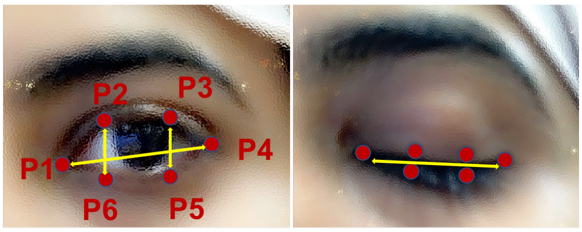


Fig. 4.3 Eye landmarks detected by Dlib.

Eq. 4.1 is used to calculate the EAR for both eyes for each frame. The average of both eyes' EAR is used. The conventional approach is to compare this averaged EAR with a particular threshold to decide whether there is an eye closure (i.e., the EAR value is lower than the given threshold) or not in each frame. For instance, Figure 4.4 shows the calculated EAR for one participant from the ZJU dataset (see Chapter 3 for details). Here, the ZJU participant has three blinks, i.e. where the calculated EAR drops below the steady state threshold of 0.25. It is clear that a single threshold could be defined for this example that would enable you to identify all the eye blinks. However, one particular threshold cannot be applied to all the participants in the ZJU dataset, as shown in Figure 4.5. This figure illustrates that each participant has a different EAR range, and therefore a threshold should be calculated based on each participant's EAR range.

$$EAR = \frac{\|p2 - p6\| + \|p3 - p5\|}{2 \cdot \|p1 - p4\|} \tag{4.1}$$

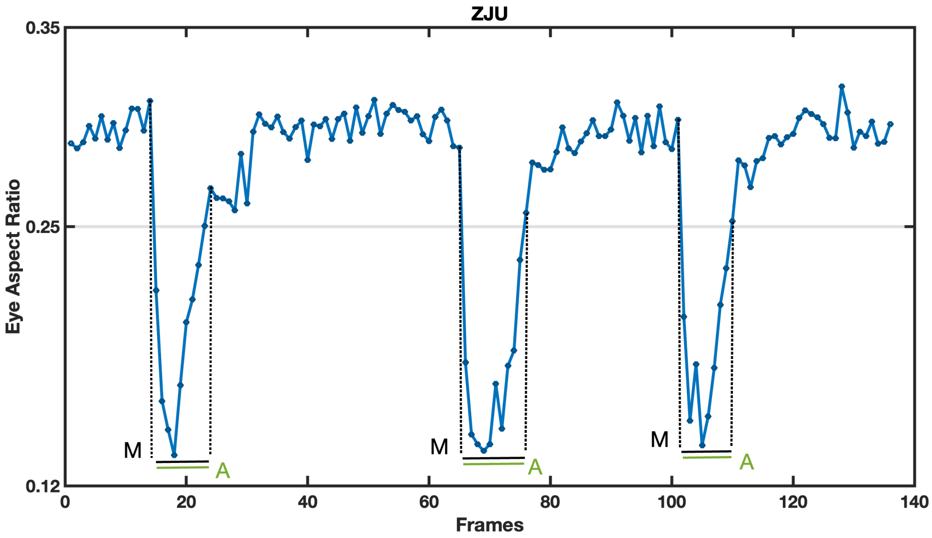


Fig. 4.4 The calculated EAR values for one recording of a participant in the ZJU data. The grey line is the threshold, the black dot and horizontal line are the manually annotated blink and the green line: is the detected true blinks (M: manually annotated blink and A: automatically detected blink).

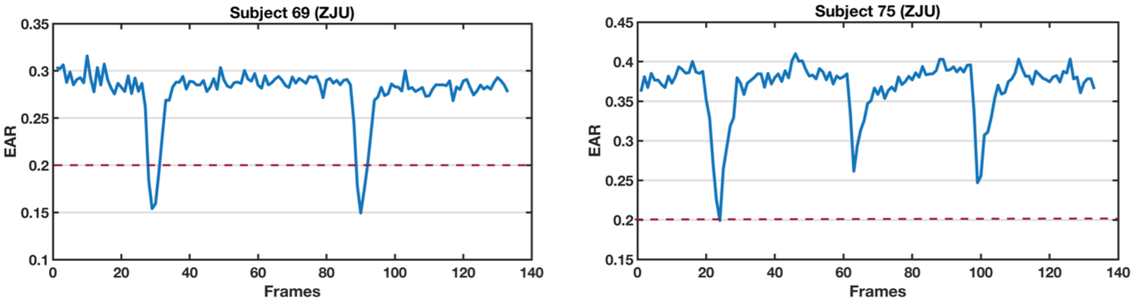


Fig. 4.5 An example of the calculated eye aspect ratio for two participants in the ZJU dataset assuming 0.2 as the threshold.

Although it would be possible to define a threshold for each of the ZJU participants, for more in-the-wild data, like the IVA_{18} dataset, this is not the case as shown in Figure 4.6 where the participant's behaviour makes it difficult to define a threshold that would work for the full file to identify eye blinks. In Figure 4.6, the location of the genuine eye blinks is indicated using a black line after observing the video manually, frame by frame. What can be clearly seen in this figure is the variability of dips in the EAR values that looks like a true blink. The grey line represents the threshold (i.e., 0.2) for determining the blink. The blink results using this threshold are indicated by green and orange lines representing the detected true and false blinks, respectively. It can be seen that four false blinks are detected in addition to the three true detected blinks.

These false blinks resulted from challenges in the dataset, such as low resolution, particularly in the eye region, low illumination, and because a monitor (showing a video animation) can be seen in the background. Another possible explanation for this is that the base mean of the EAR values does not appear to have a fixed value or a steady line as in the ZJU data. Instead, the mean base fluctuates during the session for each participant, as shown in Appendix B. Moreover, other factors, such as the blink's speed, frequency, and length, can vary significantly from one participant to another. For example, the EAR calculated for people with health conditions, such as neurodegenerative disorder (ND) and MCI, may show more false blinks due to the increase in head and body movements and tiredness, in addition to the challenges mentioned previously (Fogelton and Benesova, 2016, 2018). These challenges may increase the variation of the EAR values throughout the session. Thus, this could affect the visual-based features, such as the EBR.

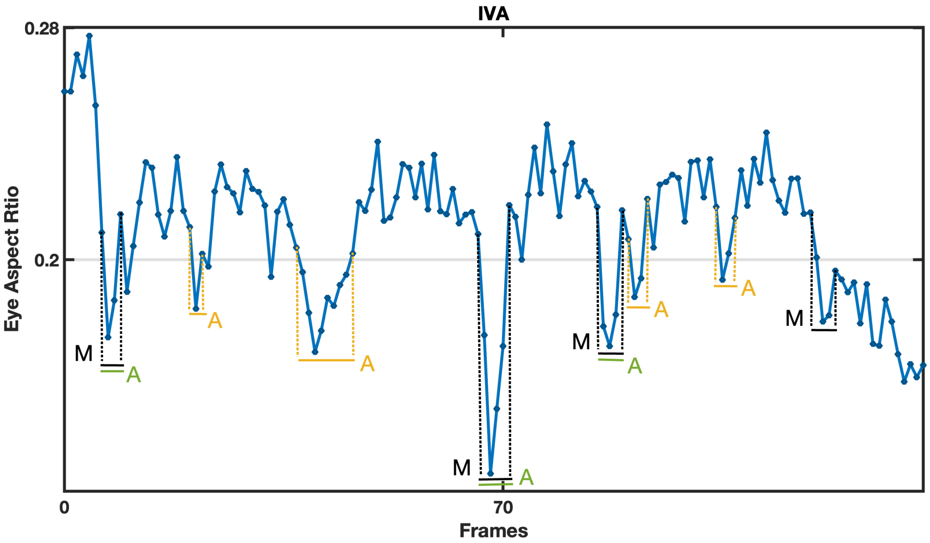


Fig. 4.6 The calculated EAR values for one recording of a participant in the IVA data. The Grey line is the threshold, the Black dot and horizontal line: are the manual annotated blink, the Orange dot and horizontal line are the detected false blink, and the Green line is the detected true blinks (M: manually annotated blink and A: automatically detected blink).

4.2.4 State Machine

Previous work has used a state machine (SM) to determine if a genuine eye blink (defined as the eye closure being longer than a certain number of consecutive frames) is detected (Alsaeedi and Wloka, 2019; Drutarovsky and Fogelton, 2014; Fogelton and Benesova, 2016, 2018). There is a range of complications that must be considered. Previous studies have given different values for the length of an eye blink – Stern et al. (1984) said that the eye blink usually lasts from 100ms to 400ms, whereas De Padova et al. (2009) gave the value as 50ms to 400ms. In addition, an eye blink may be considered incomplete when the eyes are partially closed (Portello et al., 2013) and an extended blink is between 70ms and 1s for fully closed eyes (Rodriguez et al., 2013). Also, multiple blinks may happen in the same sequence, such as double blinks, and even quadruple blinks can occur.

People with CI and related conditions may experience different behaviour regarding the EBR and blink duration. Considering that the data is in-the-wild type, determining it is not clear how long the blink duration in the SM should be. Therefore, different values for blink duration are explored as follows:

- Type 1 – one frame or any number of consecutive frames having EAR values below the threshold will be considered a blink.
- Type 2 – a sequence of two or more frames being below the threshold.
- Type 3 – a sequence of between two and 30 frames, inclusive, being below the threshold. The range corresponds to approximately 60ms to 1s due to patients who may have a long eye blink.

Figure 4.7 presents the finite SM used in this work with four states. ‘S0’ represents the initial state, while "S1" represents waiting for the next frame where the current eye status is *"Eye Open"*. ‘S2’ represents an initial eye blink detection and the blink duration constraint, which means waiting from at least n number of consecutive frames to any number of frames (NoF) while the current eye status is *‘Eye Closed’*. In this state, the previously mentioned blink duration values are explored. In Type 1, for example, the ‘S2’ will wait for at least n=1 frame or any NoF below the threshold to be considered a blink. For Type 2, the ‘S2’ will wait for at least n=2 frames or any NoF below the threshold to be considered a blink. However, Type 3 is a bit different because it will wait for at least n=2 frames or any NoF to a maximum of n=30 frames below the threshold to be detected as a blink. If the eye closure is verified based on the SM Type, ‘S2’ will transit to ‘S3’, which indicates a valid blink by changing the detected blink status to "True". In ‘S3’, if the current eye status is *‘Eye Open’*, the ‘S3’ will transit to "S1" as an indication of a new cycle of blink detection.

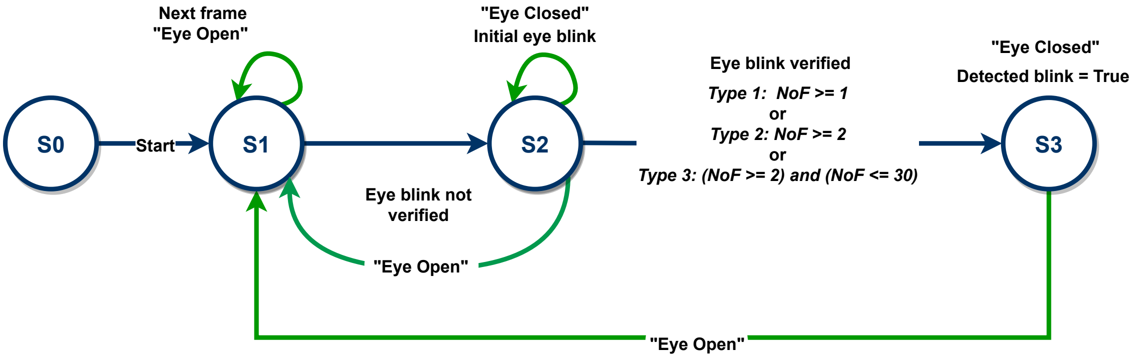


Fig. 4.7 Finite state machine to determine a true blink (NoF: Number of Consecutive Frames).

4.2.5 Automatic Calculation of a Threshold (Baseline System)

As mentioned earlier, calculating the EBR is investigated based on two approaches: i) calculating the EBR by an automatic setting of a single threshold (*baseline*) and ii) a novel approach using MTs. Each approach is examined using the two different facial landmark tracking techniques (Dlib and OpenFace) and the three SM types (1, 2 and 3).

This section presents a simple approach to constructing a baseline that depends on the mean (μ) and the standard deviation (σ) of the calculated EAR for each participant. This approach calculates a single threshold based on the participant's EAR μ and σ, so it is a participant-dependent threshold. A blink is detected when the averaged EAR for both eyes is below the μ of the EAR values minus half the σ. There is thus a separate EBR feature for the whole video for each participant. The motivation behind constructing such an approach for a baseline is to find a simple approach with good performance to compare it with the MTs approach.

In the case of the IVA_{18} dataset, the single threshold approach exhibits several issues due to the many challenges described in Chapter 3. Figure 4.8 presents the distribution of the EAR values, which are calculated using the extracted eye landmarks of OpenFace. It shows that the participants with ND or MCI have higher values of the EAR on the x-axis, which indicates head movements or turns. The sub-figure for ND participants shows that the mean of every participant on the histogram is close to the others, making the histograms overlap. The sub-figure for participants with MCI has only a small difference in the mean between the participants in the histogram. By contrast, the distribution of the EAR values for participants with functional memory disorder (FMD) shows a large difference in the mean between the participants on the histograms.

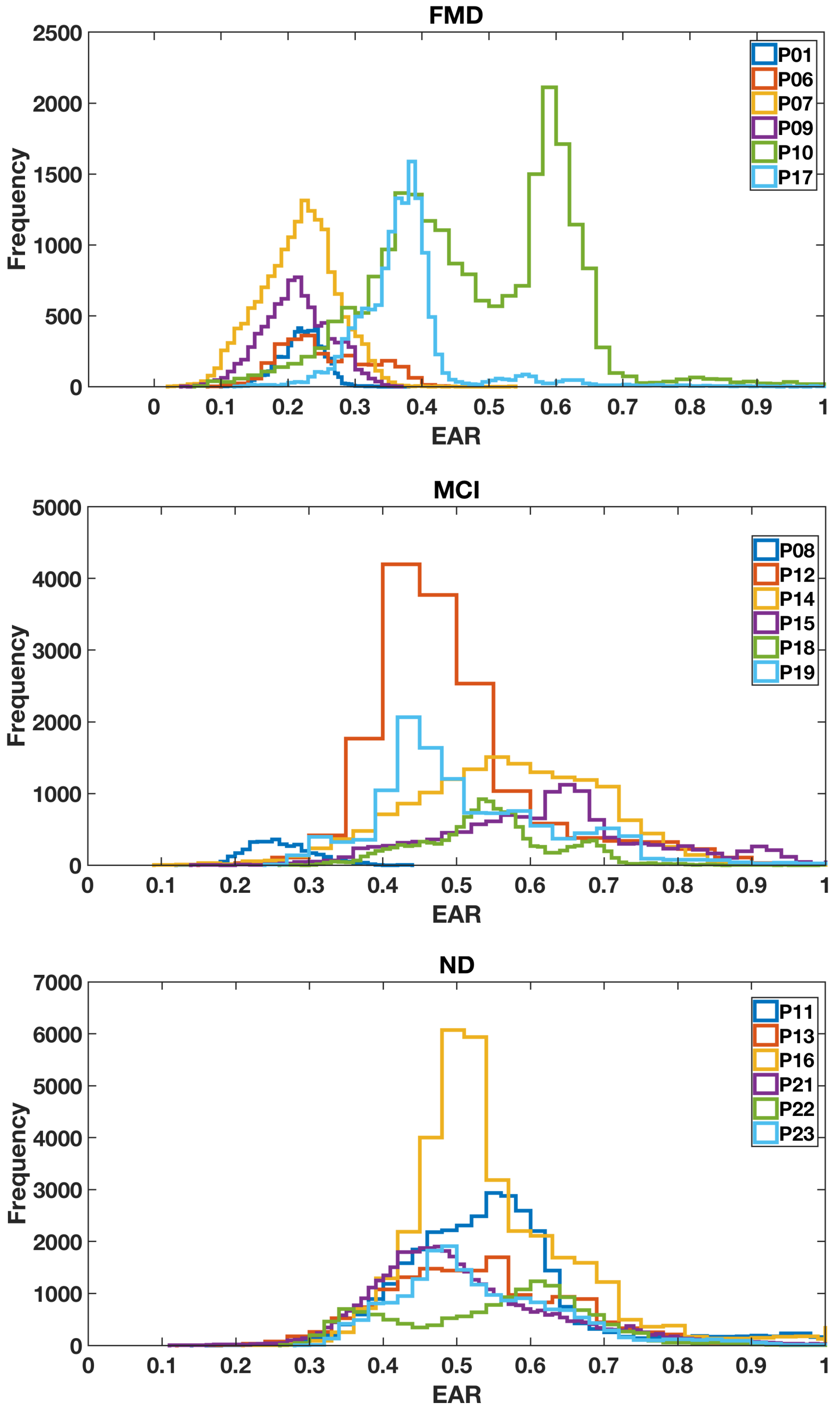


Fig. 4.8 Histograms of the computed EAR value after the extraction of the eye landmarks using the OpenFace toolkit. The X-axis of each sub-figure is cut-off at 1 to better illustrate the behaviour. There are a tiny number of EAR values up to 2.5 for FMD, 33.85 for MCI, and 13 for ND.

In addition to the variation of the EAR mean among participants (i.e., inter-speakers), intra-speaker variation can also be seen, as illustrated in Figure 4.9. This presents the calculated EAR values for the entire video of two participants in the IVA_{18} dataset. For participant P13, the EAR values range between 0.05 and 0.5 on the y-axis, whereas the values for participant P23 range between 0 and 0.65. The EAR values of both participants show a large degree of noise in the signal. These up-and-down fluctuations in signal quality make detecting the blinks very difficult, and the mean changes over time for each participant.

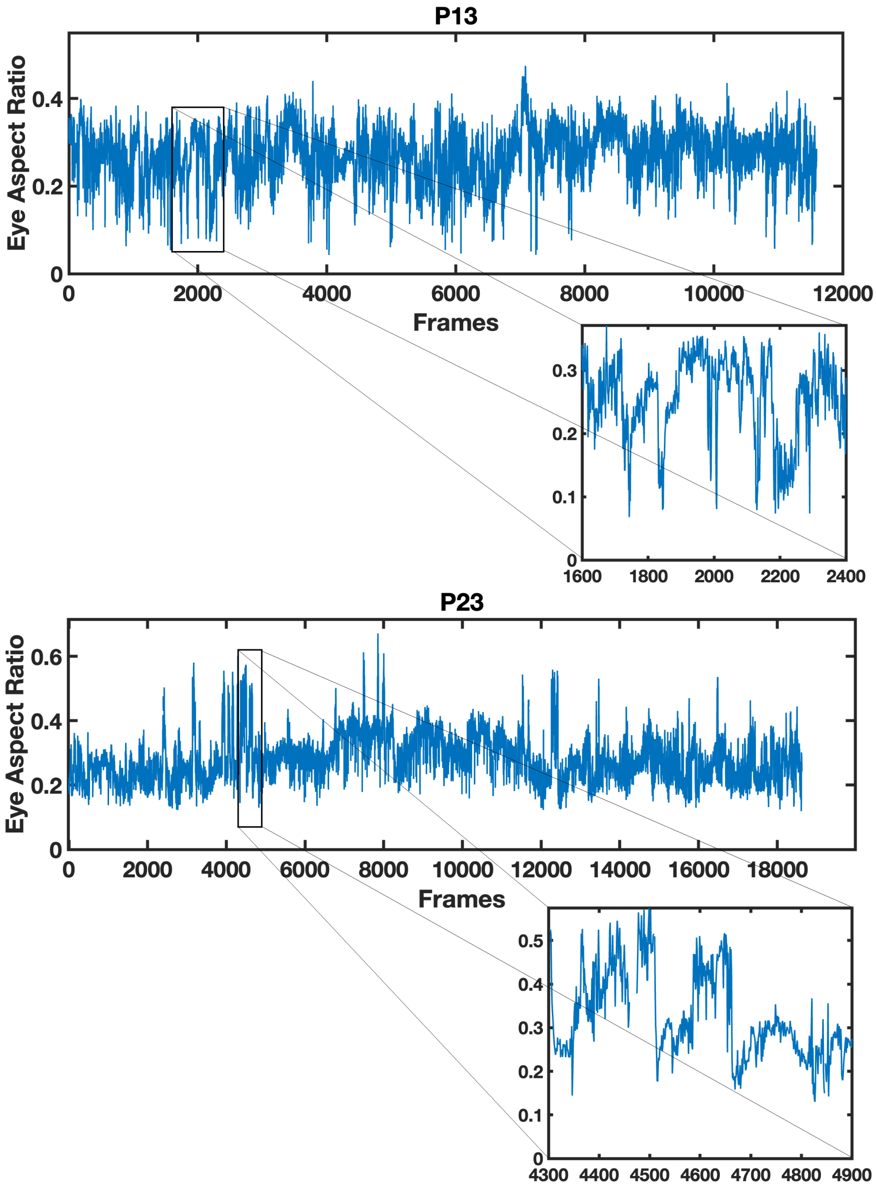


Fig. 4.9 The variations in the calculated EAR mean over the time for two randomly selected participants (P13) and (P23).

4.2.6 Multiple Thresholds Approach

Previous researchers used standard datasets to evaluate their approaches (Fogelton and Benesova, 2016, 2018; Pan et al., 2007). However, their approaches cannot detect eye blink using in-the-wild data, which includes considerable challenges (see Chapter C.22d). The dataset used in this research consists of many challenges, such as low resolution, poor illumination, participants having variable distances over time with respect to the camera, noisy background, and large head movements (see Section 3.5). This results in making the detection of the EBR considerably challenging. Another challenge is how to handle datasets for people with health conditions (e.g., MCI and ND) because they may show different spontaneous behaviours during the session, leading to more challenges that affect face detection, not just EBR detection (Taati et al., 2019). This section, therefore, proposes a novel approach that uses MTs to detect EBR. The motivation behind this approach is to address these challenges of the in-the-wild dataset and the variations that cause problems for the baseline system described in the previous section. Figure 4.10 illustrates the pipeline for the MTs approach. The pipeline involves taking the maximum and minimum EAR across all the participants to generate many thresholds of the whole video for each participant and then calculating the EBR of each threshold. The two techniques of facial landmarks detection – Dlib and OpenFace – are used to evaluate the proposed approach.

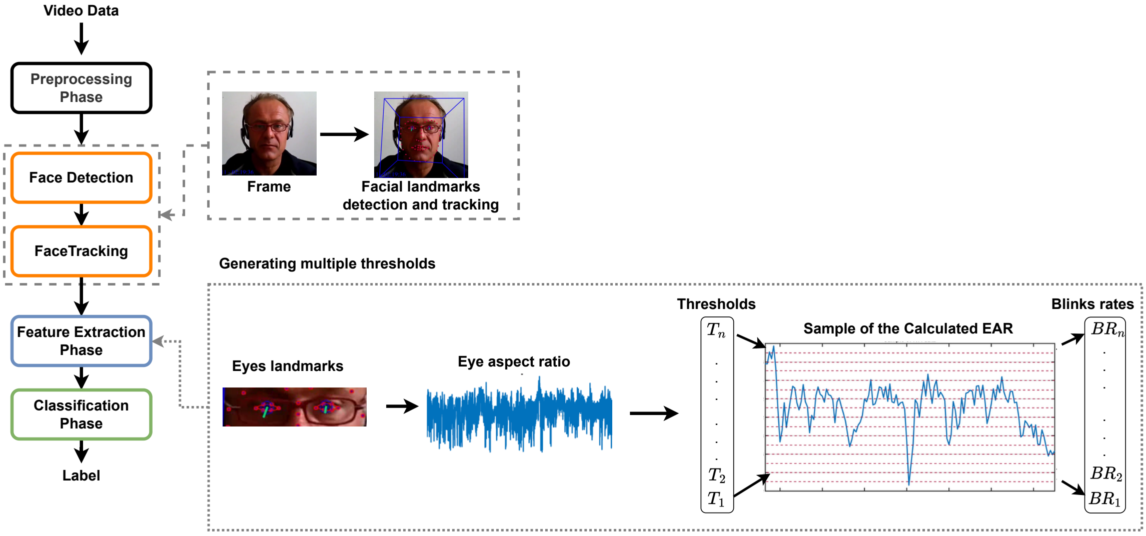


Fig. 4.10 Pipeline of the multi-threshold EBR extraction approach for each participant. Multiple thresholds (T_n) of the whole video for a participant are calculated, together with a blink rate for each threshold (BR_n).

Dlib Landmarks: MTs are generated within a particular range of the y-axis. With a minimum of 0.0 and a maximum of 0.7, a step size of 0.1 gives 0.0, 0.1, 0.2, ..., 0.6, which gives 7 thresholds. A step size of 0.01 between 0.0 and 0.7 gives 70 thresholds, and a step size of 0.001 between 0.0 and 0.7 gives 700 thresholds. These different numbers of thresholds result in 7, 70, and 700 blink rates features, respectively.

Figure 4.11 illustrates the proposed approach on the two different kinds of datasets mentioned previously in Section 4.2.5. When the MTs approach is applied to a participant in the ZJU dataset using type 2 of the SM, it can be observed that most of the thresholds give the correct number of blinks, which is three blinks. In contrast, applying the MTs approach to a 140-frame sample of a participant in the IVA_{18} dataset reveals that the number of detected blinks is mostly different for each threshold. This indicates the difficulty of detecting the true blinks from false ones and shows that using MTs could be a useful and efficient solution to capture most of the blinks in data that involve high variation in the mean of the EAR over the time of the session.

OpenFace Landmarks: MTs are calculated within the bounds of a certain range of the y-axis. This range is from 0.0 to 34. Similar to Dlib as described above, the range is divided into 7, 70 and 700 thresholds, leading to 7, 70 and 700 EBR features for each participant. Interestingly, OpenFace exhibits very high values of the EAR (e.g., 34, 12 and 25) that can be an indication of head turns or movements, occluded faces, or any issue from the previously mentioned challenges in Section 3.5.2, whereas Dlib loses the facial landmark tracking during those periods.

Another challenge in the data is the issue of EAR outlier values. Figure 4.12 shows the calculated EAR using Dlib (orange) and OpenFace (blue) for an ND participant who exhibits two extremely high values, which are indicated with a red rectangle. These abnormal values are considered outliers. It can be seen clearly that Dlib lost the facial landmark tracking for a number of frames in the red rectangle. In data mining applications, outlier detection is commonly used to detect and remove or ignore anomalous data points from the data (Tukey et al., 1977). The outliers are addressed by considering any value above $(\mu + (3 \times \sigma))$ as an outlier, where σ is the standard deviation. After the outlier calculation for each participant, the minimum outlier is 0.65, which is rounded to 0.7, making the new range for all participants to generate MTs from 0.0 to 0.7 (previously 0.0 to 34). Any number above 0.7 is then considered an outlier.

4.2.7 Classification and Evaluation Metrics

The classification involves a binary classification (ND/MCI, ND/FMD and MCI/FMD) and a three-class classification (ND/MCI/FMD), the results of which are reported in the next

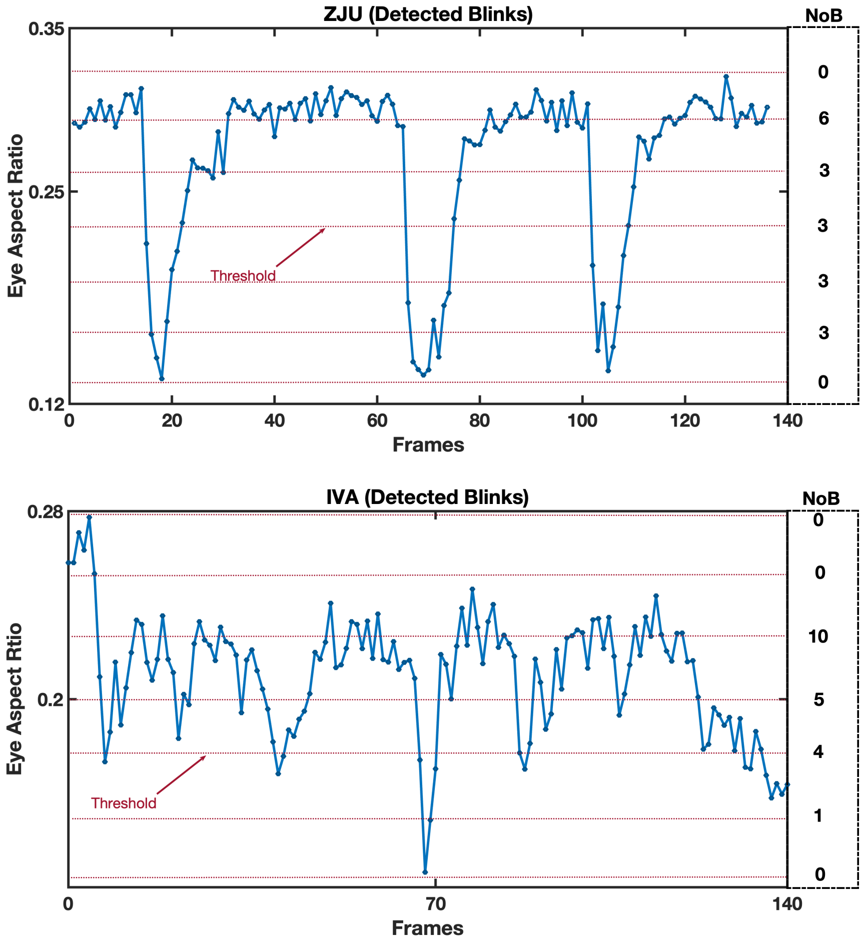


Fig. 4.11 The generated multiple thresholds for two participants: one participant from the ZJU dataset and one from the IVA_{18} dataset. Thresholds are indicated by horizontal dotted lines. The column to the right of each sub-figure gives the number of detected blinks (NoB) for each threshold.

section. First, the setting for a baseline of each classification problem is established. In this experimental work, the performance of different supervised machine learning classifiers is investigated: SVM with linear (L-SVM) and RBF (RBF-SVM) kernels, logistic regression (LR), k-nearest neighbours (kNN), and decision trees (DT). Traditional classifiers are selected instead of deep learning models due to the size of the data. The parameters are optimised for each classifier using a grid-search from Scikit-learn with held-out data to enhance the classification accuracy. Participant-independent, stratified cross-validation (CV) with 3-fold is used for each classifier. After estimating the parameters with the highest CV accuracy for each fold, they are averaged across all folds. For the SVM with Linear and RBF kernels, the regularization parameter (C) and the gamma coefficient are set to 2000 and 10, respectively. For LR parameters, the penalty multi-class and regularization are set to L2 and

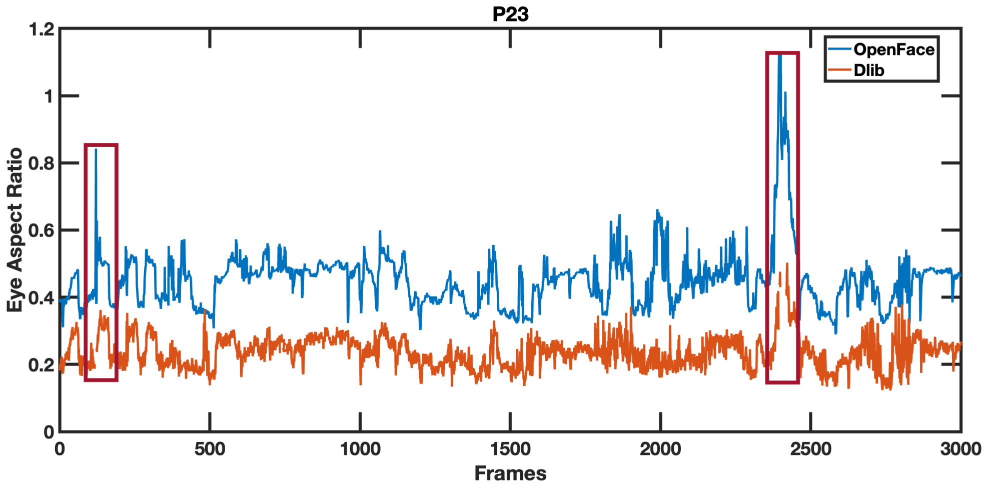


Fig. 4.12 EAR calculated for ND participant (P23) using Dlib (orange) and OpenFace (blue).

2000, respectively. For kNN, the number of neighbours is 8 with uniform as weight. For DT, the minimum numbers of splits are set to 3 and 11 for 3-class and 2-class classification, respectively.

The IVA_{18} data used in this chapter consists of 18 participants (ND: 6, MCI: 6, and FMD: 6). Thus, for evaluation, participant-independent, stratified CV with six-folds is used because CV is a common technique for evaluating machine learning models and showing the reliability of the results for small datasets (Murphy, 2012). In each fold, three participants are held out as a test set, and all the remaining participants are used in the training set. Each fold is split to maintain the sample distribution in each class. The confusion matrix shows the prediction outcomes and is usually utilised to analyse the classification results because it helps to visualise its outcomes. The performance for each classifier is the average performance across all the test sets. Since the data is balanced, the accuracy metric is used. For more information about the classifiers and evaluation metrics, see Section 2.5.

4.3 Experimental Results

4.3.1 Baseline System

The single automatic threshold approach is first evaluated on a standard dataset (ZJU). Multiple annotators, following different guidelines (e.g., ignoring any blink at the beginning or the end of the videos and counting double blinks as one blink), have annotated this dataset in order to provide a ground truth annotation of blinks, as shown in Table 4.1. For the work in this thesis, the annotation and video files are visually inspected. Fogelton and Benesova

(2016) claimed that the total number of blinks in this dataset is 261, but a further 4 blinks (at the beginnings and ends of files) are added to this annotation, resulting in a total of 265 blinks. The observed difference in the number of blinks is because the annotators of Fogelton and Benesova (2016) have excluded any blinks lying at the beginnings and ends of the video files. In this chapter, the blinks at the videos' beginnings and ends are counted as two different blinks. Moreover, double blinks are counted as two different blinks, following the approach in (Drutarovsky and Fogelton, 2014; Fogelton and Benesova, 2016). In eye blink detection, the average of both eyes is taken. For example, Radlak et al. (2015) detected eye blinks using only the right eye of the participant on the ZJU dataset.

The baseline score is calculated based on 79 of the 80 videos. One video was omitted due to the face angle of one of the participants being upward, and thus Dlib was unable to track the facial landmarks from the frames. The performance of the single threshold is measured using precision, recall, and f-measure metrics following related work (Drutarovsky and Fogelton, 2014; Fogelton and Benesova, 2016, 2018). Table 4.1 shows the results. The average f-measure is 92.5%, which means that the baseline approach can detect blinks correctly. Even though the single threshold approach is very simple, the obtained result is considered very good. A possible explanation for this might be that this dataset does not include head movement and eye occlusion.

Comparison to Related Work

Previously, researchers have used the same ZJU dataset for their evaluation. Unfortunately, often, their procedures for evaluation are either not described well, or the source code is not available to allow direct comparison. Even though their evaluation procedures are different, Table 4.1 presents an approximate comparison using precision, recall and f-measure. It can be seen that the baseline approach gave the highest recall score, which means that it is able to detect each genuine blink with a score of 100%.

Table 4.1 The performance of eye blink detection on ZJU for the baseline approach compared to related work. GT: ground-truth, DB: detected blink, TP: true positive, FP: false positive, and FN: false negative.

Study	**Precision**	**Recall**	**F-measure**	**GT**	**DB**	**TP**	**FP**	**FN**
Soukupová and Cech (2016)	99.2%	97.3%	95.2%	261	256	254	2	7
Anas et al. (2017)	98.0%	89.8%	93.7%	213	193	190	3	23
Baseline	97.8%	100%	92.5%	265	271	265	6	0

Several previous studies only used the f-measure metric to measure the performance of their experiments. Table 4.2 presents a comparison of eye blink detection performance between the baseline approach and previously published approaches on the ZJU dataset. Although the achieved performance is less than the f-measure achieved in previous work, it is still comparable to theirs because of the simplicity of the approach that does not include advanced techniques (i.e., neural networks).

Table 4.2 The performance of EBR detection on ZJU (F-measure) for the baseline approach compared to related work (WGD: weighted gradient descriptor, ZCDA: zero-crossing detection algorithm, SD: standard deviation, and NN: neural network).

Study	**F-measure**	**Approach**
Radlak and Smolka (2013)	99.2%	WGD + ZCDA
Soukupová and Cech (2016)	95.2%	SVM
Anas et al. (2017)	93.7%	NN
Fogelton and Benesova (2018)	97.6%	NN
Baseline	92.5%	Mean + SD

As stated previously, the motivation behind constructing the single threshold approach as a baseline is to find a simple approach with sufficient performance to compare it with the MTs approach. The performance of the single threshold on ZJU is 92.5%, which is considered acceptable performance to be used on the IVA_{18} data as a baseline system.

Applying the Baseline Approach to the IVA_{18} Dataset

The baseline approach is applied to the IVA_{18} dataset to differentiate the ND, MCI and FMD classes. To accomplish this, three types of SM are explored as mentioned in Section 4.2.4. The experimental results for the different classification problems using the baseline and MTs approaches are shown in Tables 4.3 and 4.4, respectively. Differentiating two groups, such as the ND/MCI and MCI/FMD, using the baseline approach gives only results close to the chance level, 58% and 50%, respectively. These two tasks are considered very challenging even in the clinic (Wakefield et al., 2018). For the three-way problem, using Dlib landmarks achieved an accuracy of 61% compared with 50% using OpenFace landmarks. It can be seen that using different types of SM does not show a significant difference in the performance of the classifiers.

4.3.2 The MTs Approach

The performance of the proposed MTs approach is investigated on multi-class classifications: ND/MCI/FMD, ND/MCI, ND/FMD and MCI/FMD, as shown in Table 4.4. The performance obtained using the MTs approach achieved the highest accuracy in the three-way classification

Table 4.3 Classification accuracy in percentages (%) when using the baseline threshold approach with different classifiers: ($Linear-SVM^1, rbf-SVM^2, kNN^3, LR^4$, and DT^5) for IVA data.

SM	Technique	ND/MCI/FMD	ND/MCI	ND/FMD	MCI/FMD
Type 1	Dlib	56^{4}	$58^{1,4}$	$75^{1,3,4}$	$83^{1,2,4}$
	OpenFace	50^{4}	75^{4}	$75^{1,2,4}$	$50^{1,4,5}$
Type 2	Dlib	61^{5}	58^{4}	83^{3}	$83^{1,2,3,4}$
	OpenFace	$50^{3,4}$	67^{4}	$75^{2,4}$	$50^{1,3,4,5}$
Type 3	Dlib	$56^{3,5}$	$50^{1,2,3,4,5}$	83^{3}	$83^{1,3,4}$
	OpenFace	50^{4}	75^{4}	$75^{1,2,3,4}$	$50^{1,3,4,5}$

compared to the baseline system using Dlib and OpenFace landmarks with 89% and 78%, respectively. For Dlib landmarks, types 2 and 3 of the SM and 70 threshold features give the best performance with an accuracy of 89%, 83%, 100%, and 92% for ND/MCI/FMD, ND/MCI, ND/FMD and MCI/FMD classes using $L-SVM^1$. These results are significantly better than the baseline results. A significant difference is found between the baseline and the MTs approach with ($p<0.05$).

For OpenFace landmarks, type 3 of the SM and 700 features obtained the highest accuracy with 72%, 100%, 92%, and 92% for ND/MCI/FMD, ND/MCI, ND/FMD and MCI/FMD classes, respectively, using DT^5. From table 4.4, it can be seen that using SD to remove the outliers helps to improve only the three-way problem with 78% accuracy using DT^5 and 7 features for each participant. The accuracy of the classifiers is significantly improved from the baseline results, and the difference between the two approaches' results is considered statistically significant with ($p<0.05$). The MTs methodology is intuitively simple, but it provides efficient results.

The results show the importance of EBR on its own as a feature to distinguish between CI types. Prior studies have utilised visual and audio-visual features for detecting CI from healthy people. For example, Tanaka et al. (2019) achieved 94% as accuracy by employing only the visual modality. In addition, Tanaka et al. (2016) and Tanaka et al. (2017) obtained 93% accuracy using SVM and 82% using LR in CI detection after combining language, speech, and visual features. Although their results are good, their works show limitations, such as limited data size, the data being recorded in a lab-controlled environment, and the participants being limited to only Japanese people. Unlike prior studies that considered ND and MCI as one group, the research described in this thesis focuses on differentiating among people with ND, MCI and FMD. Even though this work used a small dataset, the data used includes in-the-wild scenarios.

Table 4.4 Classification accuracy in percentages (%) when using the novel MTs approach using different classifiers: $(Linear-SVM^1, rbf-SVM^2, kNN^3, LR^4$, and $DT^5)$ for IVA data.

SM	Technique	Threshold No.	ND/MCI/FMD	ND/MCI	ND/FMD	MCI/FMD
Type 1	Dlib	7	$58^{1,3,4}$	58^{4}	$83^{1,4}$	83^{1}
		70	72^{1}	67^{1}	$92^{4,5}$	$\mathbf{92}^{4}$
		700	72^{1}	67^{1}	92^{4}	$83^{1,4}$
	OpenFace	7	33^{3}	50^{3}	$50^{2,3}$	$42^{1,2,3,4}$
		70	$50^{1,3,4}$	$75^{1,4}$	75^{4}	$50^{1,2,3,4}$
		700	56^{1}	$58^{2,3,5}$	83^{3}	$\mathbf{92}^{1}$
	OpenFace with SD	7	$44^{1,4}$	$58^{1,2,3}$	$83^{1,3}$	75^{5}
		70	56^{1}	58^{3}	83^{3}	$\mathbf{92}^{1}$
		700	$44^{1,2,3}$	58^{3}	83^{3}	$\mathbf{92}^{5}$
Type 2	Dlib	7	67^{4}	67^{2}	92^{1}	83^{1}
		70	$\mathbf{89}^{1}$	83^{1}	$\mathbf{100}^{1,5}$	$\mathbf{92}^{1,4}$
		700	78^{4}	$75^{4,5}$	$\mathbf{100}^{5}$	$\mathbf{92}^{4}$
	OpenFace	7	33^{3}	$50^{3,4}$	$50^{2,3,4}$	$42^{1,3,4}$
		70	$44^{1,2,4}$	$67^{1,3,4}$	$75^{1,3,4}$	$50^{1,2,3,4}$
		700	50^{3}	$50^{3,4,5}$	$83^{1,4}$	$67^{1,3,5}$
	OpenFace with SD	7	$50^{3,5}$	$50^{3,4}$	$83^{2,4}$	67^{5}
		70	$50^{3,4}$	58^{5}	83^{24}	$\mathbf{92}^{5}$
		700	61^{4}	50^{3}	83^{2}	$\mathbf{92}^{5}$
Type 3	Dlib	70	$\mathbf{89}^{1}$	83^{1}	$\mathbf{100}^{1,5}$	$\mathbf{92}^{1,4}$
	OpenFace	700	72^{5}	$\mathbf{100}^{5}$	92^{1}	$\mathbf{92}^{5}$
	OpenFace with SD	7	78^{5}	67^{4}	$83^{1,2}$	$\mathbf{92}^{1}$

4.3.3 Analysis

As explained above, this study compares the performance of two different facial landmark tracking techniques to handle data recorded in the wild. The findings of this work show that the performance of CI detection was better when using Dlib landmarks than OpenFace landmarks. This result can be explained by the fact that Dlib only detects frontal and semi-frontal frames and removes noisy frames, whereas OpenFace detects all the frames. As previously mentioned, the high values of the EAR could be an indication of head movements, turns, or any one of the challenges described in Chapter 3. The results of this study show that removing those frames using an outlier detection approach (OpenFace with SD) improves only the three-way classification problem, which is still less than the achieved accuracy using Dlib.

In order to understand more about the performance of the classifiers, a confusion matrix is used. Figure 4.13 shows that when using both techniques (Dlib and Openface), ND and

FMD are predicted correctly. However, some MCI participants appeared to be predicted as ND when OpenFace was used. These results indicate that MCI is the most challenging class to classify from ND in two-way and three-way classification problems. A possible explanation for this might be that people with MCI experience an increase in EBR related to memory problems, and this increase may indicate a risk of progression from MCI to ND, which is in line with the results of a previous study (Ladas et al., 2014).

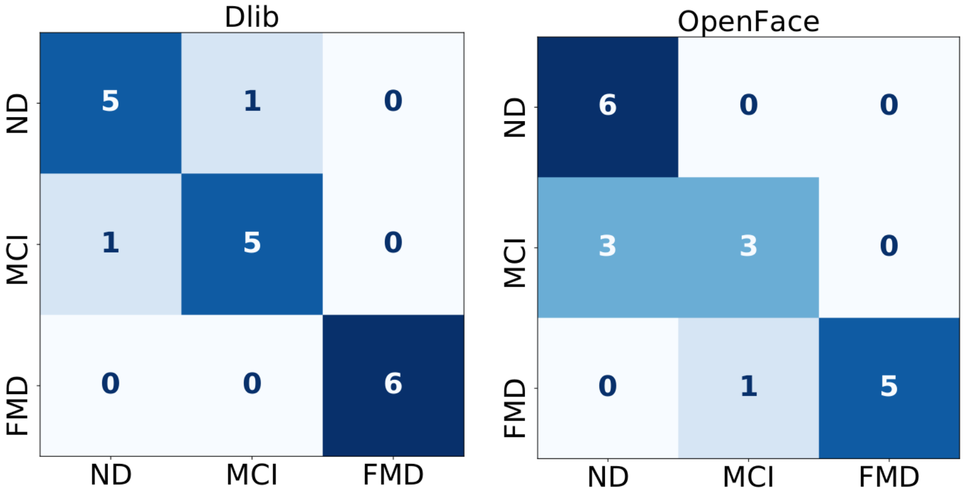


Fig. 4.13 Confusion matrix for the three-way classification (ND vs. MCI vs. FMD) using DLib and OpenFace (rows: true labels and columns=predicted labels).

4.4 Summary

This chapter has described a research that was undertaken to detect dementia and related conditions by designing an approach for calculating EBR using in-the-wild data. Given the challenge of data recorded in the wild, experiments were carried out to investigate: 1) the possibility of calculating EBR using this kind of data and 2) the feasibility of this approach for automatically classifying CI and related conditions.

This research has presented a novel MTs approach for EBR data that could be used for CI detection. This investigation used two different facial landmark techniques – Dlib and OpenFace – to explore their performance on the kind of data recorded in the wild. The results confirmed that Dlib landmarks gave better results than Openface landmarks, especially in three-way classification, with an accuracy of 89% using Dlib and 78% using OpenFace. The findings have shown the significance of EBR as a cue to distinguish ND, MCI, and FMD from each other. The reason behind these results is that Dlib only detects facial landmarks when the face is frontal or semi-frontal, otherwise, it loses track of these landmarks, which is the opposite of OpenFace, which detects all the frames. According to these results, we can infer that using the frames in which the face is facing the camera could give better performance than also including the noisy frames.

The most important limitation of this study lies in the fact that the size of the dataset was limited. This issue is common in any study that involves human-participant data (Tanaka et al., 2017, 2016). However, collecting human participants with CI and related conditions is more challenging than healthy individuals' data. Further work needs to be carried out in order to validate the obtained results on a larger dataset, which will be described in the following chapter. After that, more visual features will be investigated for the IVA_{18} dataset and then evaluated on a larger dataset.

Chapter 5

Exploring the Robustness of the MTs Approach on the IVA_{52} Dataset

> *"The most beautiful people we have known are those who have known defeat, known suffering, known struggle, known loss, and have found their way out of the depths. These persons have an appreciation, sensitivity, and understanding of life that fills them with compassion, gentleness, and deep loving concern. Beautiful people do not just happen."*
>
> *— Elisabeth Kübler-Ross*

5.1 Introduction

The previous chapter developed a robust multiple thresholds (MTs) approach to extract the eye blink rate (EBR) and overcome the issues involved in in-the-wild data (IVA_{18}), as described in Section 3.5. This chapter evaluates the methodology of the EBR feature extraction on a larger dataset, referred to as IVA_{52}. This dataset consists of both the IVA_{18} dataset, which was used previously in Chapter 4, and the IVA_{34} dataset, which includes new additional recordings of participants with neurodegenerative disorder (ND), mild cognitive impairment (MCI), functional memory disorder (FMD) and healthy controls (HC), described in Section 3.5. The IVA_{34} dataset varies in a number of ways from IVA_{18}. In particular, it is recorded in a number of different environments, such as clinics and homes, using different devices (e.g., laptops and smartphones). These two datasets are combined due to the small number of participants in the three groups in the IVA_{34}. This evaluation, therefore, involves assessing the extent to which the performance of this methodology is affected by increasing the data size and variations in the recording environments and the devices used.

The participants' behaviour in the IVA_{34} dataset is first analysed and then compared with participants' behaviour in the IVA_{18} dataset in terms of the EBR feature. This chapter also examines additional classification tasks between different groups and combinations of groups in order to allow direct comparison with related work (Tanaka et al., 2019, 2017, 2016).

The chapter is organised into six main sections. Section 5.2 covers the analysis of the calculated eye aspect ratio (EAR). Section 5.3 investigates the use of the MTs approach on the IVA_{52} dataset. Section 5.4 shows the results achieved using the MTs approach. Section 5.5 presents the different improvements to the MTs approach to address the increased variation in the IVA_{52} dataset. Section 5.6 demonstrates the results of several experiments on the MTs approach and for different classification tasks. Finally, Section 5.7 contains a discussion and a conclusion.

5.2 Analysing the Calculated EAR

In this chapter, the IVA_{52} dataset is used, which is described in detail in Chapter 3. Prior to the facial landmarks tracking, the dataset is pre-processed, as described in Chapter 4. Then, the EAR is calculated for each frame using the eyes' landmarks. This section provides an analysis of the participants' behaviour in the new IVA_{34} dataset in terms of the EAR and then compares it with the IVA_{18} dataset.

5.2.1 General Behaviour of the IVA_{34} Participants

Chapter 3 outlined the main in-the-wild conditions for the IVA_{34} dataset and compared the difference in the conditions when the videos were recorded at home or in the clinic. Figure 5.1 shows the EAR calculated for some participants in the IVA_{34} who recorded their session in the clinic and at home using laptop and smartphone devices. The rest of the participants' figures are presented in Appendix C. The figure shows extremely high values, which are mainly the result of the variations in the recording environments (e.g., clinic and home), the variation in the device used (e.g., laptop and smartphone), and many challenges independent of the recording environment, such as the participants' behaviour during the session, which are explained in Chapter 3. It can be seen that the bottom figure, which represents a video recorded at home using a smartphone, shows many high values compared with the upper figures due to issues related to the use of this kind of device. An example issue is participants holding their phones at a lower angle from their face, which makes their eyes look partly closed. This difference may affect the calculation of the EAR from those calculated using laptop-recorded data. These challenges lead to false calculations in the eyes' landmarks and loss of tracking of the facial landmarks.

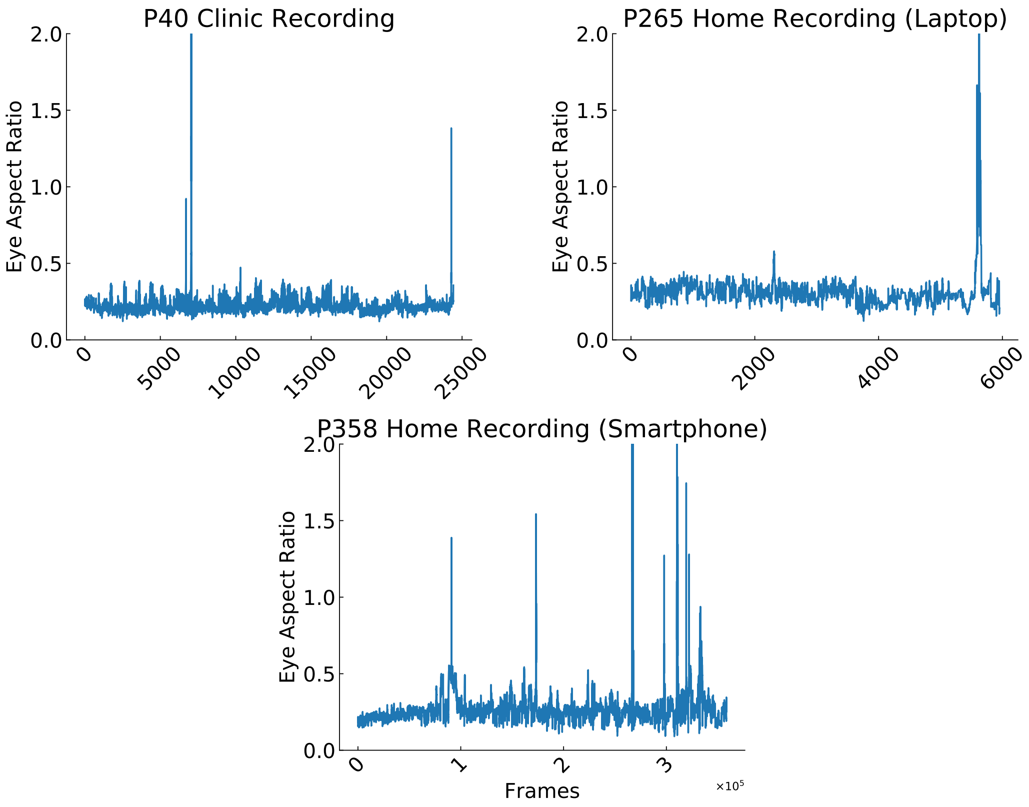


Fig. 5.1 An example of the calculated EAR values for different participants who made the recording in a clinic and at home.

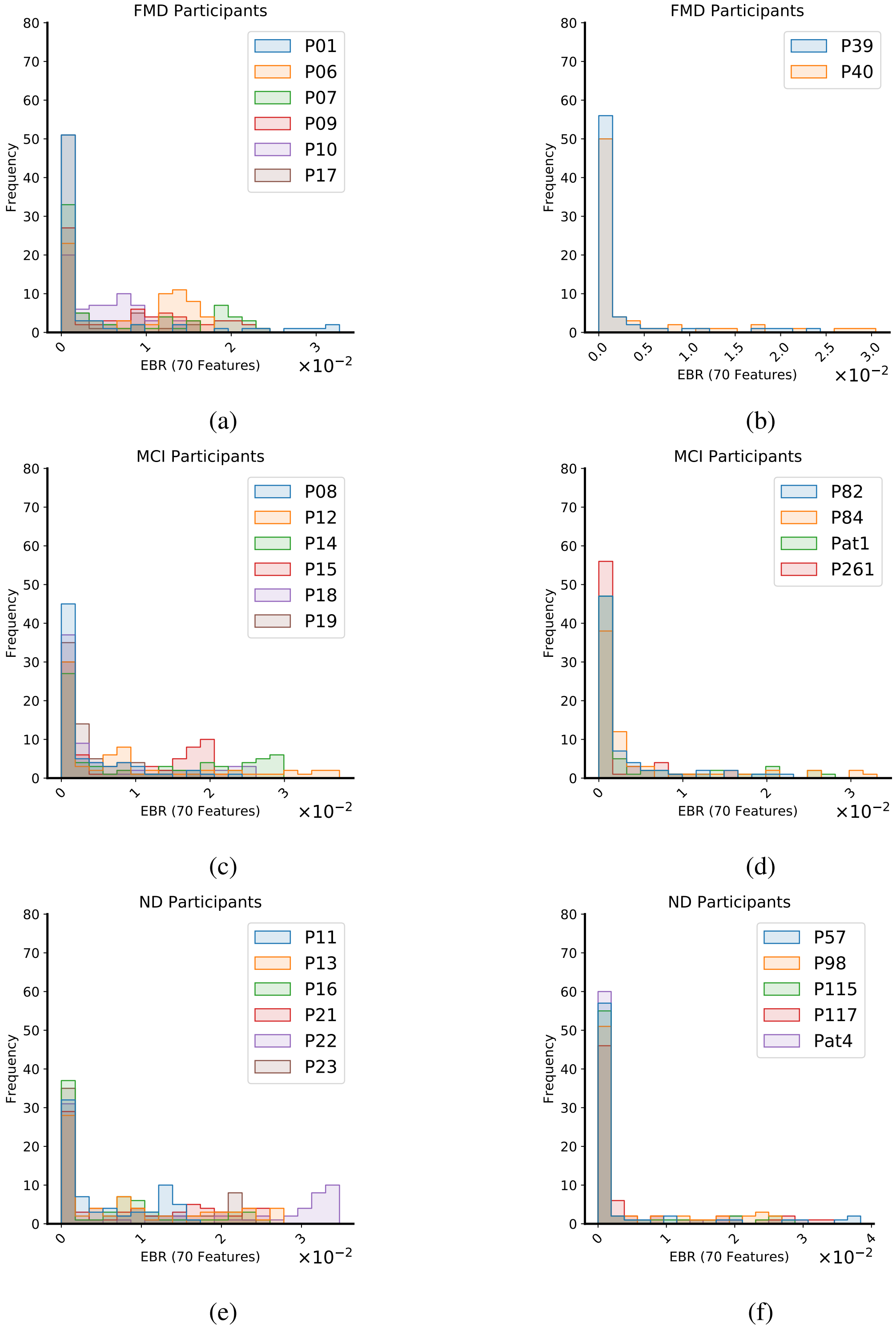
5.2.2 Comparing IVA_{34} with IVA_{18}

To analyse the IVA_{34} further and understand the difference between this dataset and the IVA_{18} dataset, the EBR feature is calculated for each participant based on the EAR. The results are shown in Figures 5.2 and 5.4 as a histogram presenting the EBR feature on the x-axis and the frequency on the y-axis. These figures show all the participants across the two datasets (IVA_{18} on the left and IVA_{34} on the right) for each of the diagnostic labels: ND, MCI, FMD and HC.

Figure 5.2 shows that the three groups of the IVA_{34} dataset (right column) have a similar pattern to the extracted EBR features as the HC group in Figure 5.4. That is to say, most of the features are skewed to the left and lie on the zero value on the x-axis, indicating that their EAR range is small. For instance, the calculated EAR values of participants P39 and P40 are shown in Figure 5.3 and can be compared with the extracted EBR features in

Figure 5.2. Figure 5.3 shows that the range of the EAR is very small. This indicates that the eyes of the participants appear very small in the video in which the EAR is calculated. When the MTs approach is applied to them, based on the minimum and maximum over all participants (minimum = 0 and maximum = 20), there would be thresholds resulting in zero or very close to zero values because their EAR ranges from 0.21 to 0.39 and 0.20 to above 2.0. Consequently, only two to three thresholds would give meaningful results for the EBR values that are not zero or close to zero when the number of thresholds is 7. However, when the number of thresholds is 70 or 700, the number of features with zero or very close to zero values is going to increase.

Comparing the different diagnostic classes, generally, the FMD class shows the highest number of zeros, whereas ND shows the lowest number of zeros. This may have been caused by the participants' movements during the session, which causes a lot of noise in the signal. The MCI shows more zero values than ND but fewer zero values than FMD. In terms of comparing IVA_{34} with IVA_{18}, the IVA_{34} participants P39 and P40 show higher frequency values on zero or close to zero from participants of IVA_{18} with the same class. One reason for this is suggested above, which is related to the observed small eye regions of the participant, and another reason is the low resolution (poor video quality). P01 from IVA_{18} displays the same issues (low resolution, small eye region), which is why high counts are seen on zero value on the x-axis of the histogram. In addition, having high counts on zero or close to zero from MCI participants of IVA_{34} more than IVA_{18} with the same class is observed. The same observation is seen for the ND class from both IVA_{18} and IVA_{34}. These observations indicate that the range of EAR values for the IVA_{34} participants is smaller than the IVA_{18} data.



(a)

(b)

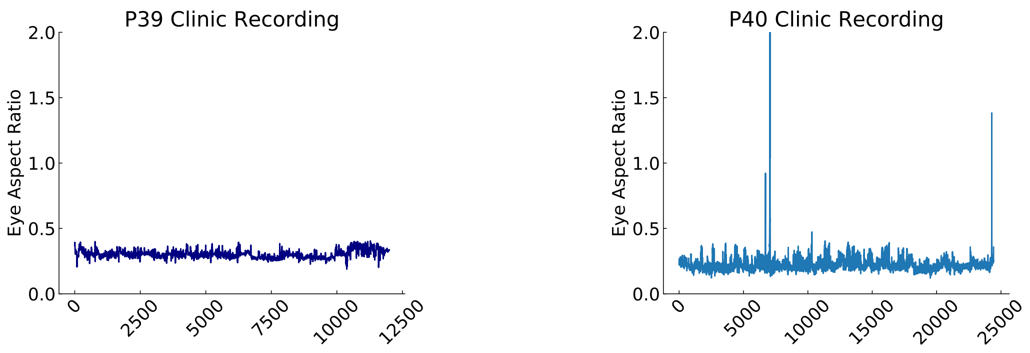
(c)

(d)

(e)

(f)

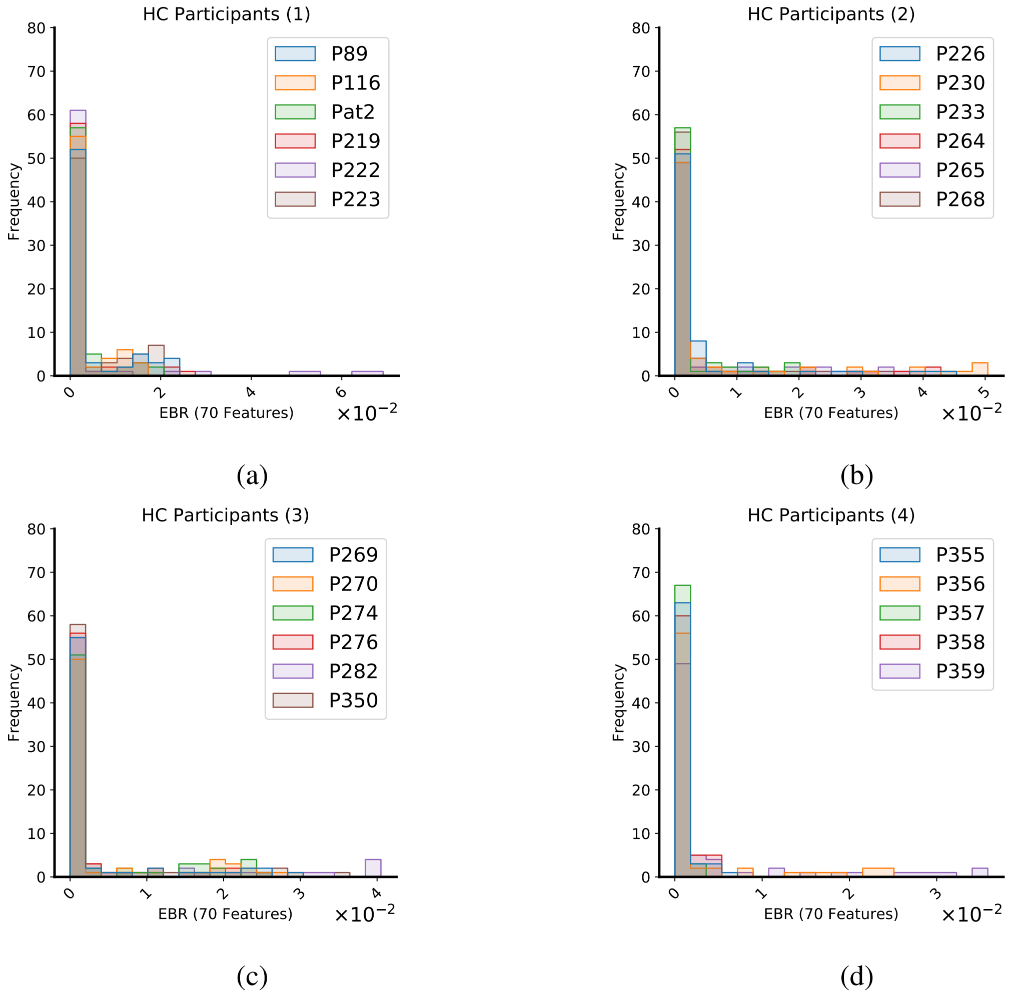
Fig. 5.2 Histogram plots show the data distribution based on the class from IVA_{18} (left column) and IVA_{34} (right column) data recorded in different recording environments using laptop and smartphone (Part 1).



(a)

(b)

Fig. 5.3 The calculated EAR for two participants with FMD from IVA_{34}. The horizontal lines in the plots represent the calculated thresholds for these participants.



(a)

(b)

(c)

(d)

Fig. 5.4 Histogram plots showing the data distribution of healthy control group from IVA_{34} data recorded in different recording environments using laptop and smartphone (Part 2).

Figure 5.5 shows the mean (blue) and the 3rd SD (orange) of each participant. Notably, the mean and the SD of participants in IVA_{18} are higher than in IVA_{34}, which confirms the previous observations that the IVA_{34} participants have a lower range of the EAR than the participants in IVA_{18}. Taking all the above evidence together, the proposed approach that worked for IVA_{18} (see Chapter 4) cannot achieve a high-performance on IVA_{34} due to the new challenges. This requires improvements to the MTs approach to overcome these challenges.

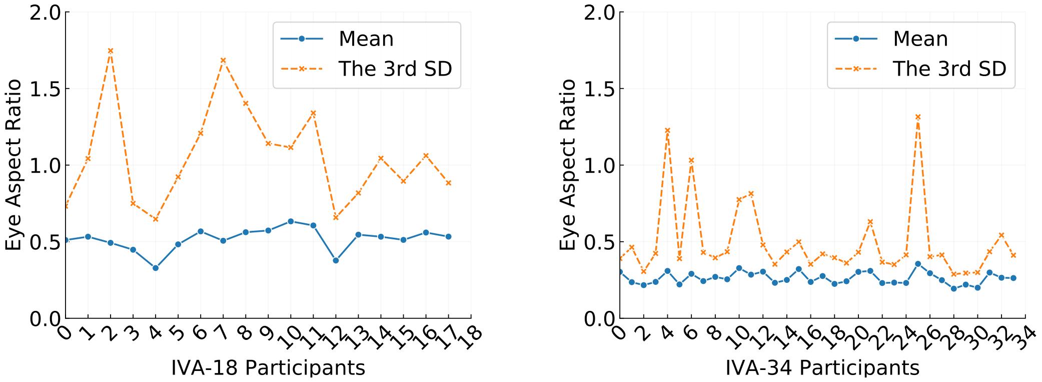


Fig. 5.5 The mean and the 3rd SD of every participant in both datasets, IVA_{18} (left) and IVA_{34} (right).

5.3 Investigating the MTs Approach

This section focuses on evaluating the MTs approach on a larger dataset. As stated in Chapter 4, the MTs approach generates many thresholds using the minimum and maximum EAR over all the participants (OAP). However, in this chapter, the definition of MTs is evolved to become participant-dependent (PD) due to the issues mentioned above related to the calculation of the EAR in the IVA_{34} dataset. It calculates the thresholds based on the minimum and maximum for each participant, which will be explained later in Section 5.5. Throughout this chapter, the abbreviation MTs-OAP will be used to refer to multiple thresholds using the OAP approach and the abbreviation MTs-PD will be used to refer to multiple thresholds using the PD approach.

The IVA_{34} data includes a small number of participants in the three categories, 2 participants with FMD, 4 with MCI, 5 with ND and the rest are HC. Therefore, the IVA_{34} data is combined with IVA_{18} to experiment with a large dataset (IVA_{52}). Three main sets of experiments are carried out to explore the performance of the MTs-OAP approach using the combined dataset IVA_{52}. Firstly, a three-way (ND, MCI and FMD) classification task is conducted on the IVA_{52} data to compare its results directly with those of the IVA_{18} dataset results. Then, two-way classification tasks are conducted by dividing the data into an HC group and memory problems (MP) group, which are the ND, MCI and FMD classes. A final set of experiments divides the data into a dementia (D) group, containing the ND and MCI classes, and a Non-D group, including the FMD and HC classes.

5.4 Experiment: MTs-OAP Approach

This section describes the testing of the performance of the MTs-OAP approach in classifying memory-related problems (MP) and HC from each other. A discussion of the purpose, data, classifiers used and results obtained is then presented.

5.4.1 Feature Extraction and Classifiers

In this chapter, the MTs-OAP approach with type 2 of the state machine (SM), which is described in Chapter 4, is used to detect the EBR to enable a faster run-time and the many experiments needed to test the system performance on the MTs-OAP approach. Firstly, the MTs-OAP is tested in a three-way classification (ND, MCI, and FMD). The same four classifiers as were used when testing IVA_{18} performance in the previous chapter are used again here: SVM with linear kernel, logistic regression (LR), k-nearest neighbour (KNN) with Uniform weight, and decision trees (DT). Using the same classification task and the same classifiers allows a fair and direct comparison between the IVA_{18} and the combined data IVA_{52} results. The IVA_{18} data includes 6 with ND, 6 with MCI, and 6 with FMD, and the combined data IVA_{52} includes 11 with ND, 10 with MCI, 8 with FMD and 23 HC. The classifiers are trained using the Python Scikit-learn package. The classification is participant-independent-stratified k-fold cross-validation. Some hyper-parameter values are optimised using a grid search, and the rest are set to their default values. For each classifier, four metrics are computed, accuracy, recall, precision and F-measure. Details of the calculation of these metrics can be found in Section 2.5.

5.4.2 Results

Table 5.1 and Figure 5.6 present the classification results for the IVA_{18} and the combined data IVA_{52} with DT. The results are above the chance level and show that the IVA_{18} results outperform the IVA_{52} dataset results. The IVA_{18} gave the best results using 700 thresholds with 61% accuracy, while the combined data needed a smaller number of thresholds to achieve the best performance with 52% accuracy. The reason behind this result can be the imbalanced data in the three groups. From the confusion matrix, the combined data helps to improve only the prediction of the ND class, whereas the predictions of the FMD and MCI are reduced. Most of the incorrectly classified labels are for MCI, mostly as ND. This shows that the MCI group can be challenging to distinguish from the ND group, even in the clinic. This may indicate an issue related to the MTs-OAP approach. To confirm this assumption, two more experiments are conducted.

Table 5.1 Classification results for the three-way problem when the MTs-OAP approach with SD is used on IVA_{52} and compared to the results obtained on IVA_{18} using type 2 of the SM.

Data	**No. of thresholds**	**Accuracy**	**Precision**	**Recall**	**F-measure**
IVA_{18}	700	**61%**	**63%**	**61%**	**61%**
IVA_{52}	70	52%	49%	51%	47%

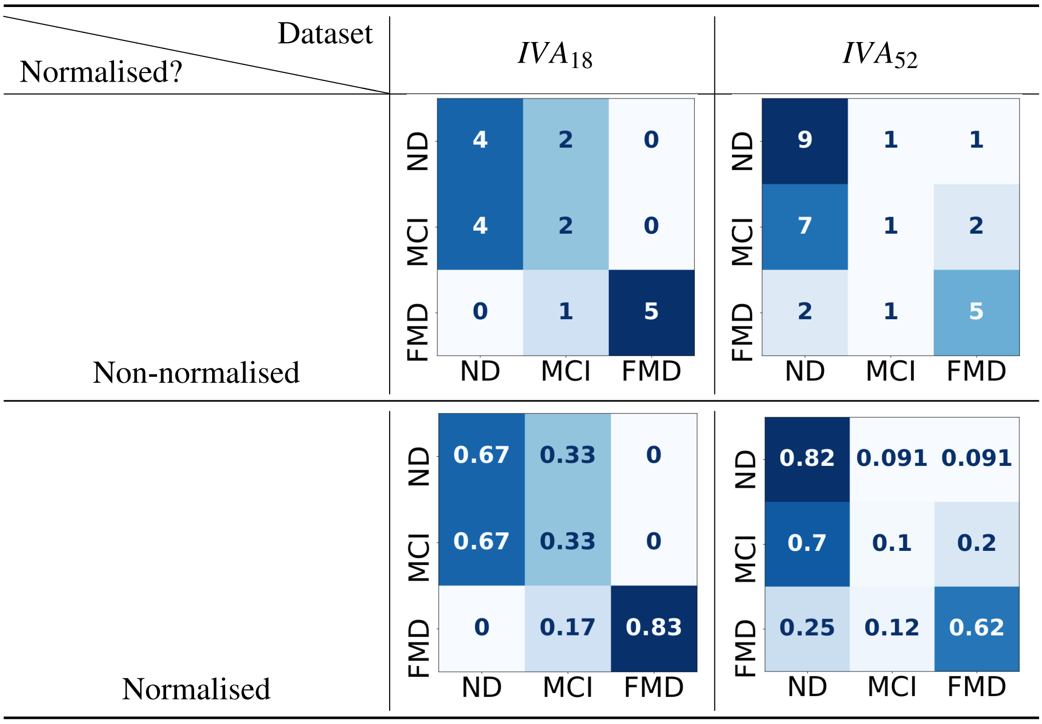
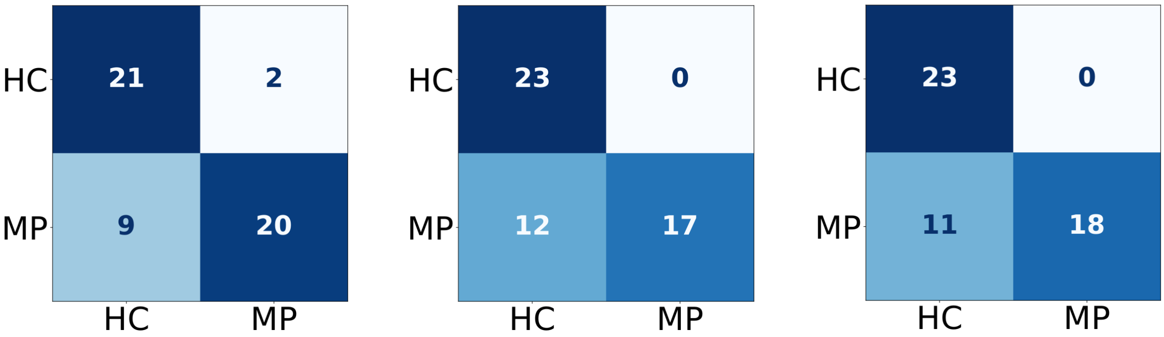


Fig. 5.6 Confusion matrix for the three-way classification between ND, MCI and FMD using type 2 of the SM and 70 thresholds to compare the predictions between IVA_{18} in (a) and IVA_{52} in (b). (rows: true labels and columns: classified labels).

Next, two two-way classification tasks are investigated since the IVA_{52} data includes many HC participants. The first two-way classification task involves dividing the data into two groups: HC and MP. The experiment is carried out on a different number of thresholds, and the highest results are obtained using the KNN, as shown in Table 5.2 and depicted in Figure 5.7. The metrics in the table may, at first glance, be considered very good. However, when these results are analysed using the confusion matrix shown in Figure 5.7, they show that the incorrectly classified participants from the MP group are all new participants from the IVA_{34} dataset (2 FMD, 4 MCI, and 5 ND).

Table 5.2 Classification results for the two-way problem (HC vs. MP) using the MTs-OAP approach on IVA_{52} with type 2 of the SM (OAP:over all the participants, MP: includes ND, MCI and FMD).

No. of thresholds	Accuracy	Precision	Recall	F-measure
7	**80%**	80%	80%	**79%**
70	79%	83%	79%	77%
700	**80%**	**84%**	**81%**	**79%**



(a) (b) (c)

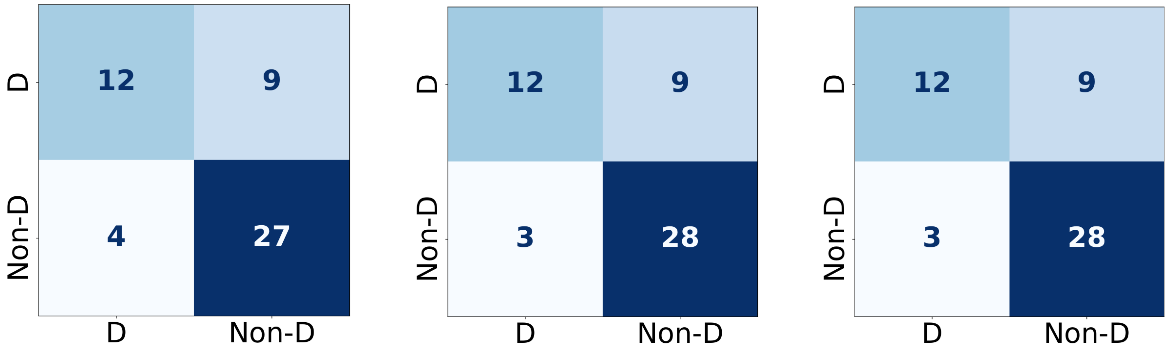
Fig. 5.7 Confusion matrix for the two-way classification between HC and MP using a different number of features or thresholds (a) 7 features, (b) 70 features, and (c) 700 features. (rows: true labels and columns: classified labels)

The second two-way classification task investigated is performed to classify people with D from Non-D people. The data is divided into D and Non-D. The D group consists of two groups: MCI:10 and ND:11. The Non-D group includes HC:23 and FMD:8. Table 5.3 and Figure 5.8 show the classification results for the different numbers of thresholds. The KNN classifier achieves the highest scores in all metrics using 70 and 700 features. Like the first sub-experiment, the incorrectly classified labels of the D are from the new data (4 with MCI and 5 with ND). However, the misclassified labels from the HC group are from the IVA_{18} data (3 with FMD).

These two experiments confirm the assumption from the three-way classification task result and the previously mentioned observation when the mean and the SD of the EAR for each participant in IVA_{18} and IVA_{34} individually were calculated and exhibited a difference between the two datasets. The difference between them is that the EAR's mean and SD for the IVA_{18} participants are higher than those of the IVA_{34} participants, and that is why participants with MP are misclassified every time as HC. The low mean and SD of the IVA_{34} participants' EAR are a result of small pixels occupying the eye region and some challenges mentioned above. Consequently, several issues appeared, such as 1) the prediction of the MCI class from the three-way classification and 2) the incorrect prediction for all the new data from the two-way classifications. For that reason, the MTs approach needs to be improved to overcome the new challenges from the new data. The next section explores and discusses the improvements in the MTs approach.

Table 5.3 Classification results of the two-way problem (Non-D vs. D) using the MTs-OAP approach on IVA_{52} with type 2 of the SM.

No. of thresholds	**Accuracy**	**Precision**	**Recall**	**F-measure**
7	75%	75%	72%	73%
70	**78%**	**79%**	**74%**	**75%**
700	**78%**	**79%**	**74%**	**75%**



(a) (b) (c)

Fig. 5.8 Confusion matrix for the two-way classification between D and Non-D using a different number of features or thresholds, (a) 7 features, (b) 70 features and (c) 700 features. (rows: true labels, columns: classified labels, D: ND/MCI and Non-D: HC/FMD)

5.5 Investigating the MTs-PD Approach

The MTs-PD approach described above is developed to overcome the new challenges by addressing the challenges of having a very small range of EAR values and making the threshold calculation PD rather than taking the minimum and maximum EAR OAP. As stated previously, the IVA_{52} data includes many challenges that resulted in many extremely high values in the EAR. In addition, calculating the mean and the 3rd SD of each participant using the EAR showed a significant difference between the IVA_{18} and the IVA_{34} datasets. Another approach for determining the upper boundary (UB) is, therefore, explored: the interquartile range (IQR), which will be described in the following section. A comparison between the UB of the IQR approach and the 3rd SD approach is shown in Figure 5.9 for the individual participants in the IVA_{18} and IVA_{34} datasets.

Considering the above-mentioned problems and the variation in the recording environments and devices used in the dataset, three experiment setups are conducted, as shown in Figure 5.10. Experiment setup 1 measures the effect of the SD and IQR approaches in removing the outliers. Then, the performance is measured on two-way and three-way classification tasks using the IVA_{52} dataset, as in the previous section. Experiment setup 2 measures the effect of the variability according to the recording environment. This involves splitting the HC group into different sets based on the challenges, starting with the videos that are close to a clinic recording, then videos recorded at home using a laptop with more challenges and ending with videos recorded at home using smartphones. The performance is measured on two-way classification tasks: MP vs. HC and D vs. non-D. Finally, experiment setup 3 is carried out on a combination of two-way and four-way classification tasks.

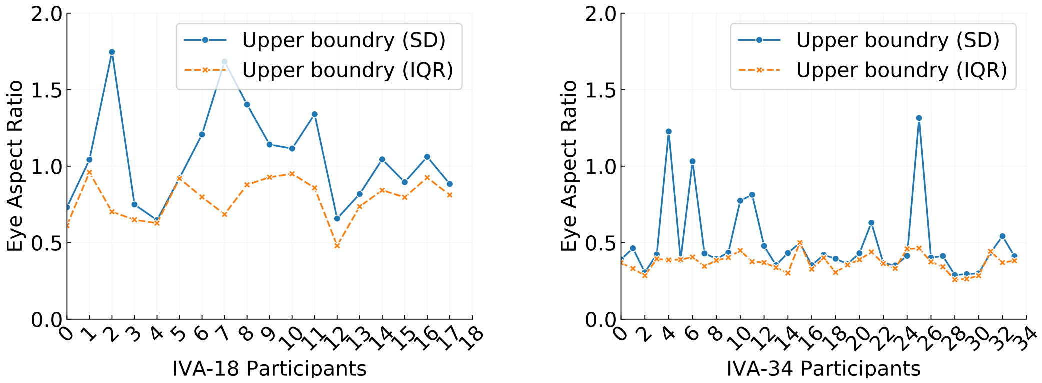


Fig. 5.9 Comparison between the 3rd SD and the UB of the IQR using both of them on both datasets; IVA_{18} (left) and IVA_{34} (right).

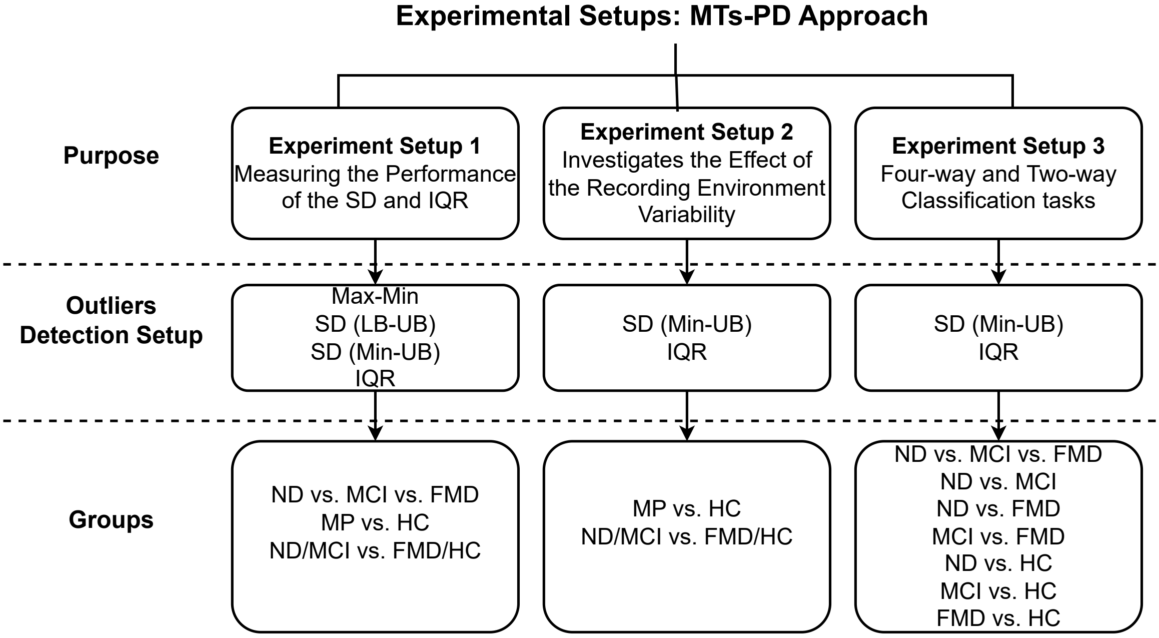


Fig. 5.10 The three different experimental setups using the MTs-PD approach.

5.6 Experiment: MTs-PD Approach

5.6.1 Feature Extraction and Classifiers

The MTs-PD approach is investigated using three different setups to calculate the thresholds, which result in different patterns of the EBR feature. These different setups are described in the following section with the results obtained. The performance of the same four classifiers and metrics used in the previous experiment are utilised again. Only one classifier result is shown here to enable a direct comparison with the previous experimental setup. The hyper-parameters are optimised using a grid search; the rest are set to their default values. Participant-independent-stratified K-fold cross-validation is used.

5.6.2 Results

Experiments are carried out based on different setups: Min-Max, SD and IQR, which are explained as follows:

- **Min-Max setup:** This takes the minimum and maximum values of the EAR for each participant. However, taking the whole signal into account does not show encouraging results, as shown in Tables 5.4, 5.5, and 5.6, even in the IVA_{18} results presented in Chapter 4. Hence, an approach to determine the appropriate data UB is required to ignore the extreme values as outliers. The most common approaches to finding the outliers, which are used in several previous studies are based on the standard deviation (SD) (Yang et al., 2018) and the IQR (Rousseeuw and Croux, 1993). These approaches are used here to find the UB of this data.

- **SD (LB-UB) setup:** this calculates the lower boundary (LB) and UB values of the EAR as mean +/- three times the SD (Simmons et al., 2011). The general equations for calculating the threshold are presented in Equations 5.1 and 5.2. Other studies used more aggressive choices by using mean +/- 2 or 2.5 times SD, indicating an outlier level of 0.62% and 2.28%.

$$LB = Mean - a * SD \tag{5.1}$$

$$UB = Mean + a * SD \tag{5.2}$$

  The mean and the SD are calculated to determine the outlier values and a is a control parameter determined by the user. When the value is small, more values will be included in the range of the LB and UB. The most common value of a is 3 because the number of outliers is expected to be small (Yang et al., 2019). This approach also did not show good results, as can be seen in Tables 5.4, because the LB calculation included values that were lower than the needed range. For example, if the minimum value of EAR for a particular participant equals 0.2, the LB value for this participant, taking the mean into account, could be equal to or less than 0.0, which will generate thresholds that do not include useful information for calculating, and their EBR values are zeros.

- **SD (Min-UB) setup:** this only differs from the previous set-up in calculating the LB as the minimum EAR value. This change helps to give a better performance in the two-way classification in Table 5.5 that distinguishes people with MP from HC. However, Table 5.4 shows no improvements in the results obtained from the three-way classification task, whereas the performance in classifying the D class from the Non-D class is reduced, as shown in Table 5.6. Therefore, another approach to determine the UB is employed to enhance the results.

- **IQR setup:** this is defined as the difference between the 75th and 25th percentiles of the data. These values are called Q1 and Q3, respectively. The IQR approach is used because it is very robust in detecting outliers and is not sensitive to high values. The IQR equation to compute the threshold is:

$$IQR = Q3 - Q1 \tag{5.3}$$

$$UB = Q3 + b * IQR \tag{5.4}$$

The most common value of *b* is 1.5 (Simmons et al., 2011). When the performance is tested using the IQR to determine the UB, the three-way classification results are improved, as can be seen in Table 5.4. In addition, the two-way classification that differentiates people with ND and MCI from HC and FMD shows better performance than using SD (see Table 5.6). However, the performance in classifying HC from MP decreases compared to the performance obtained using the SD.

It was assumed that IQR would be the feasible solution that would give the highest performance because the approach is insensitive to extreme values and could capture only the useful information from the EAR values and result in a very good performance. However, the results contradict this assumption for some classification tasks. In fact, it appears that extreme values play a key role in the classification. In other words, the extreme values may show a pattern that helps to differentiate these classes from each other more than when the range only includes the main signal. More investigation is carried out for these different setups because the IQR is not concluded to be a feasible method to use for the next step. Accordingly, two sub-experiments are performed to analyse the previous results and determine which method for determining the UB could be used for the different classification tasks. Experiment setup 1 measures the performance of the SD and IQR approaches by varying their factors *a* and *b*. Experiment setup 2 measures the effect of the variability of the recording environment by reducing the number of participants in the HC group to include only the good data for which recording settings are close to the IVA_{18} data.

Experiment Setup 1: Measuring the Performance of the SD and IQR Approaches

In this experiment, a range of factors for both approaches are explored to show how they will perform. The varied range of the factor for both approaches, SD and IQR, is investigated via three classification tasks: a three-way problem (ND vs. MCI vs. FMD) and two two-way problems (HC vs. MP) and (HC/FMD vs. ND/MCI). The range is from 1 to 3 and is increased by increments of 0.5 (e.g., 1, 1.5, 2, 2.5, 3). Four metrics are calculated from each classification task, and the same classifier used previously for each classification task is again used in this experiment to enable a fair and direct comparison.

Table 5.7 presents the results of varying the factor range of the SD and IQR approaches in the three-way classification. The results indicate that using IQR to determine the UB helps to improve the system's performance better than SD. However, SD shows no improvement by varying the factor, which is reflected in the confusion matrices shown in Figure 5.11. The highest results are obtained when factor *b* equals 1.5, 2.5 and 3. The difference between the

Table 5.4 Classification results of the three-way problem (ND vs. MCI vs. FMD) when the MTs-PD approach is applied to IVA_{52} with type 2 of the SM and linear support vector machine (NoT: number of thresholds).

NoT	Approach	Accuracy	Precision	Recall	F-Measure
7	Min-Max	31%	31%	31%	35%
	SD (LB-UB)	38%	33%	38%	36%
	SD (Min-UB)	38%	33%	38%	36%
	IQR	**49%**	**64%**	**54%**	**47%**
70	Min-Max	**53%**	**51%**	**52%**	**51%**
	SD (LB-UB)	45%	45%	45%	45%
	SD (Min-UB)	44%	49%	47%	49%
	IQR	46%	47%	48%	47%
700	Min-Max	52%	51%	52%	51%
	SD (LB-UB)	**55%**	**55%**	**55%**	**55%**
	SD (Min-UB)	44%	47%	45%	49%
	IQR	49%	50%	52%	51%

Table 5.5 Classification results of the two-way problem (HC vs. MP) when the MTs-PD approach is applied to IVA_{52} with type 2 of the SM and linear support vector machine (NoT: number of thresholds).

NoT	Approach	Accuracy	Precision	Recall	F-Measure
7	Min-Max	69%	69%	69%	68%
	SD (LB-UB)	**75%**	**76%**	**75%**	**74%**
	SD (Min-UB)	74%	71%	71%	71%
	IQR	66%	66%	65%	65%
70	Min-Max	69%	69%	69%	69%
	SD (LB-UB)	74%	74%	71%	71%
	SD (Min-UB)	**78%**	**76%**	**75%**	**75%**
	IQR	68%	67%	67%	67%
700	Min-Max	69%	69%	69%	68%
	SD (LB-UB)	73%	74%	71%	71%
	SD (Min-UB)	**76%**	**73%**	**73%**	**73%**
	IQR	68%	67%	67%	67%

IQR and SD results is significant. Table 5.9 and Figure 5.13 also show that the IQR approach has a better performance than the SD approach, especially at factor $b = 2$, but no significant difference in the results between the IQR and SD.

Table 5.6 Classification results of the two-way problem (D vs. Non-D) when the MTs-PD approach is applied to IVA_{52} with type 2 of the SM and linear support vector machine (NoT: number of thresholds).

NoT	Approach	Accuracy	Precision	Recall	F-Measure
7	Min-MAx	63%	63%	62%	62%
	SD (LB-UB)	63%	67%	65%	60%
	SD (Min-UB)	58%	60%	60%	75%
	IQR	58%	60%	60%	75%
70	Min-MAx	74%	74%	73%	71%
	SD (LB-UB)	71%	76%	71%	67%
	SD (Min-UB)	64%	65%	65%	63%
	IQR	71%	73%	73%	72%
700	Min-MAx	78%	77%	77%	77%
	SD (LB-UB)	79%	79%	79%	79%
	SD (Min-UB)	64%	65%	65%	63%
	IQR	67%	67%	67%	67%

On the other hand, the classification task HC vs. MP shows that the SD approach gives a better performance than IQR for all the variations, which is the opposite result to that of the two classification task mentioned above (see Table 5.8 and Figure 5.12). From the table and the confusion matrix, it can be seen that the SD gives the highest performance when it equals 1.5. The difference between the results obtained using SD and IQR is considered to be statistically significant.

From the results obtained above, more analysis is needed by investigating the confusion matrices for one of the previous classification tasks to understand the difference between the performances obtained. Consequently, the HC vs. MP classification task is chosen because it includes all groups, and there is a significant difference between the IQR and SD results even when their factors are varied.

Figure 5.12 shows that the common misclassified labels are not from a specific dataset (i.e., IVA_{18} or IVA_{34}). The incorrect prediction between MP and HC is due to the variation in the recording environments and the devices used. These variations lead to high variations in the range of the EAR calculations that could affect the detection of participants with MP from HC. Interestingly, the HC participants are misclassified when the factor of the SD increases, while the MP participants are correctly classified. For instance, MP participants are classified incorrectly when the factor of the SD is a=1, although HC participants are classified correctly.

Table 5.7 Classification results of the three-way problem (ND vs. MCI vs. FMD) using a range of values for SD and IQR to find the factor with the highest performance score. These approaches are tested on 70 thresholds.

Factor	Approach	Accuracy	Precision	Recall	F-Measure
1	SD (Min-UB)	44%	31%	40%	32%
	IQR	**49%**	**51%**	**51%**	**51%**
1.5	SD (Min-UB)	44%	31%	40%	32%
	IQR	**53%**	**58%**	**53%**	**52%**
2	SD (Min-UB)	44%	31%	40%	32%
	IQR	**50%**	**51%**	**48%**	**45%**
2.5	SD (Min-UB)	44%	31%	40%	32%
	IQR	**53%**	**58%**	**53%**	**52%**
3	SD (Min-UB)	44%	31%	40%	32%
	IQR	**53%**	**58%**	**53%**	**52%**

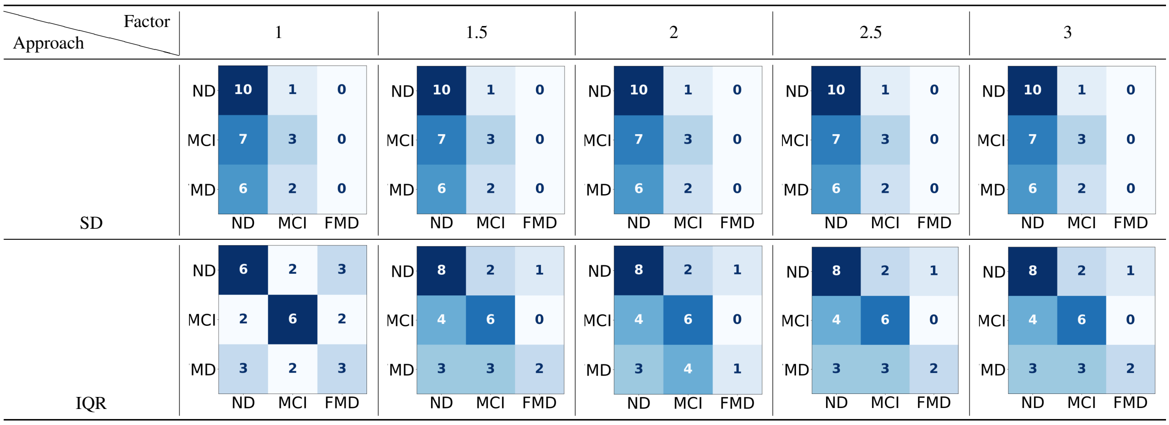


Fig. 5.11 The confusion matrices for the three-way classification (ND vs. MCI vs. FMD) using two approaches to detect the UB with a range of factors applied to 70 thresholds (rows: true labels and columns: classified labels).

Experiment Setup 2: Investigates the Effect of Variability in the Recording-Environments

In this experiment, an investigation is carried out to explore the effect of including a subset of HC participants in a clinic-like environment and increasing the number of HC participants by adding more data to include more in-the-wild data. The HC group is divided into 5 sets: set 1 includes 9 participants of HC who were recorded at a clinic or at home that is close in terms of the recording environment to clinic recordings, set 2 includes an additional 5 HC participants who used a laptop, set 3 is the combination of set 2 with 4 more HC participants

Table 5.8 Classification results of the two-way problem (HC vs. MP) using a range of values for SD and IQR to find the factor with the highest performance score. These approaches are tested on 70 thresholds or features using KNN with uniform as weight(MP= includes ND, MCI and FMD).

Factor	Approach	Accuracy	Precision	Recall	F-Measure
1	SD (Min-UB)	**77%**	**78%**	**77%**	**75%**
	IQR	63%	61%	61%	61%
1.5	SD (Min-UB)	**81%**	**80%**	**80%**	**79%**
	IQR	65%	63%	63%	63%
2	SD (Min-UB)	**78%**	**76%**	**76%**	**75%**
	IQR	69%	67%	67%	67%
2.5	SD (Min-UB)	**78%**	**76%**	**76%**	**75%**
	IQR	68%	67%	67%	67%
3	SD (Min-UB)	**78%**	**76%**	**76%**	**75%**
	IQR	68%	68%	67%	67%

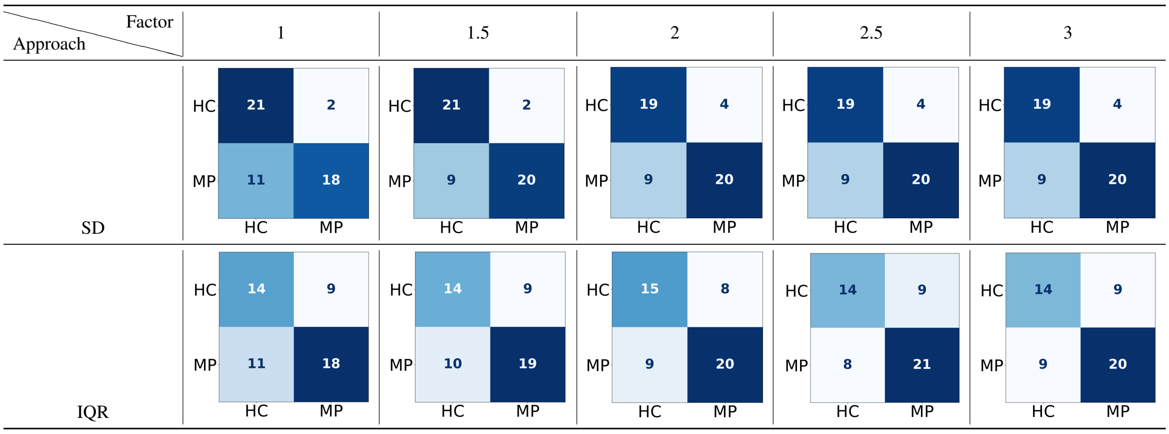


Fig. 5.12 The confusion matrices for the two-way classification (HC vs. MP) using two approaches to detect the UB with a range of factors applied to 70 features or thresholds (MP: includes ND, MCI and FMD, rows: true labels, columns: classified labels).

who used a laptop and set 4 consists of 5 HC participants who used a smartphone. To test the performance, the SD approach uses two factors: default $a = 3$ and the one that achieved the highest performance from previous results $a = 1.5$.

From the previous confusion matrix analysis, one of the clinic's participants (Pat2) is classified incorrectly for all of the SD and IQR factors variations in both of the two-way classification tasks: HC vs. MP and D vs. Non-D. Hence, this participant is removed, resulting in nine participants of HC. However, participant P89 is misclassified only for all the IQR factors in the D vs. Non-D classification task. As mentioned above, the different environments and devices for recording can affect the EAR range between the

Table 5.9 Classification results of the two-way problem (D vs. Non-D) using a range of values for SD and IQR to find the factor with the highest performance score. These approaches are tested on 70 thresholds or features using KNN with uniform as weight.

Factor	Approach	Accuracy	Precision	Recall	F-Measure
1	SD (Min-UB)	**63%**	62%	59%	58%
	IQR	62%	**64%**	**63%**	**63%**
1.5	SD (Min-UB)	67%	66%	63%	63%
	IQR	**71%**	**72%**	**71%**	**71%**
2	SD (Min-UB)	64%	64%	62%	62%
	IQR	**74%**	**74%**	**73%**	**73%**
2.5	SD (Min-UB)	67%	66%	65%	65%
	IQR	**71%**	**72%**	**71%**	**71%**
3	SD (Min-UB)	64%	64%	61%	61%
	IQR	**69%**	**70%**	**68%**	**69%**

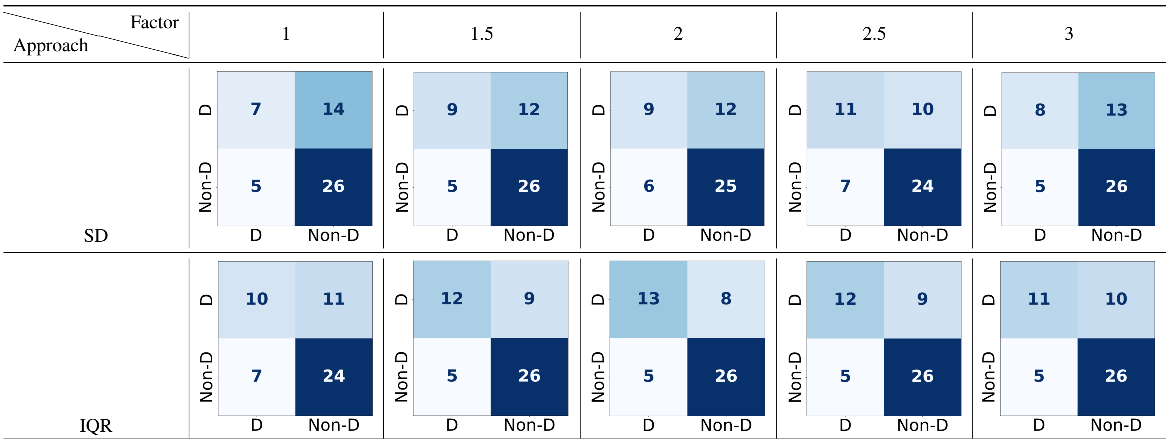


Fig. 5.13 The confusion matrices for the two-way classification D vs. Non-D using two approaches to detect the UB with a range of factors applied to 70 features or thresholds (rows: true labels and columns: classified labels).

participants with MP (e.g., ND, MCI and FMD) from IVA_{18} and IVA_{34}. Participant Pat2 is misclassified due to his behaviour that is observed in the EAR values that is close to the ND group in the IVA_{34} dataset, while P89's behaviour is close to the ND and MCI groups and misclassified in only one classification task. From the analysis above, nine HC participants whose recordings have a clinic-like environment to use with the clinic recordings for MP are chosen as set 1 in order to find out whether such recordings for HC participants can make a difference in improving the classification results or not.

Table 5.10 and Figure 5.14 show that using all the HC participants leads to a decrease in the performance of detecting participants with MP (e.g., ND, MCI, and FMD). However,

Table 5.10 Classification results of the two-way problem (HC vs. MP) using two factors for SD a= 3, the default value, and a = 1.5, the factor with the highest performance. The classification is done on 9 participants, and every time increases, the number of HC by combining more data (MP= includes ND, MCI and FMD groups; set 1= 9 participants of HC who were recorded at the clinic or at home in conditions close to clinic recordings; set 2= with an additional 5 participants with HC who used a laptop; set 3= is the combination of set 2 with 4 more participants with HC who used a laptop; set 4= 5 participants of HC who used a smartphone).

Factor	HC Number	Accuracy	Precision	Recall	F-measure
3	Set1	**85%**	**78%**	**78%**	**78%**
	Set1+Set2	79%	76%	75%	76%
	Set1+Set3	73%	72%	73%	72%
	Set1+Set4	82%	79%	79%	79%
	All HC participants	78%	76%	75%	75%
1.5	Set1	**82%**	75%	80%	77%
	Set1+Set2	77%	76%	79%	76%
	Set1+Set3	75%	75%	76%	74%
	Set1+Set4	**82%**	**79%**	**83%**	**80%**
	All HC participants	81%	82%	79%	79%

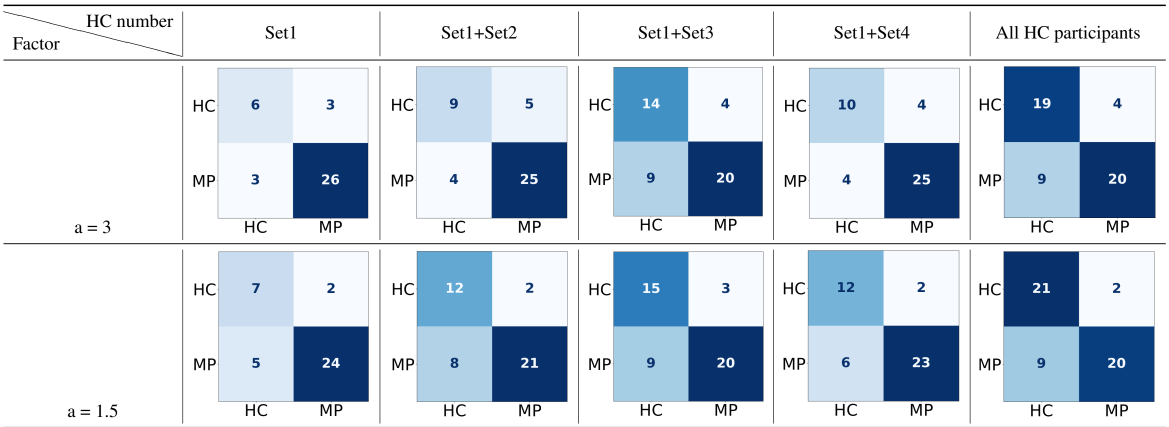


Fig. 5.14 The confusion matrices for the two-way classification HC vs. MP using two factors for the SD to detect the UB with different number of HC participants and uses 70 features or thresholds (MP: includes ND, MCI and FMD, rows: true labels, columns: classified labels).

when the number of HC participants is reduced to 9 participants, the performance of the system is improved in detecting MP, whereas the detection of HC decreases, as shown in the confusion matrix. This shows that the variations in the recording environments could affect the performance of the system. The results also show that using SD with factor 3 improves the detection of MP more, whereas using a low factor value of 1.5 enhances the prediction of

HC from MP. These results confirm the previous observation from analysing the confusion matrix of the HC vs. MP classification task.

Adding 5 more HC participants (Set1+Set2) improves the HC classification from MP with accuracy from 64% to 78% using $a = 3$. In contrast, adding all the HC participants (Set1+Set3) who recorded the session using the laptop caused a decline in the performance of the system of predicting MP from HC in percentage from 86% to 69% using factor $a = 3$. When the 9 participants of HC are combined with 5 participants who did the session using smartphones (Set1+Set4), it gives better performance of the system even than using only Set 1.

A comparison of Figure 5.12 with Figure 5.14 shows that several participants are classified incorrectly even with reducing the number of HC; P57, P117, and P21 from the ND group; P84 from the MCI group; and P01, P07, P17, and P39 from the FMD group. Participants P117, P01 and P17 are misclassified in every division of the data and factor except for factor a = 3. Moreover, HC participants are classified incorrectly when factor a is greater than 1.5.

A third sub-experiment setup is performed on a four-way and a combination of two-way classification problems.

Experiment Setup 3: Four-way and Two-way Classification Tasks

The previously described approaches are evaluated in two experiments: four-way classification (ND vs. MCI vs. FMD vs. HC) and two-way classification (ND vs. MCI, ND vs. FMD, MCI vs. FMD, ND vs. HC, MCI vs. HC, FMD vs. HC). These classification tasks are investigated using both the SD and IQR approaches, with the default factor for each one and the factors with the highest performance. The four-way classification results of using both the SD and IQR approaches with their default factors and the ones that give the highest performance are shown in Table 5.11. From Table 5.11, it can be seen that the IQR achieves the highest performance, and that changing the factor shows no difference in the obtained results. It is observed that the detection of participants with MP (e.g., ND, MCI, and FMD) is better when the IQR is used. In contrast, the performance decreases when the SD is used, especially for FMD participants classified as HC and MCI. This observation confirms the conclusion reached in the previous chapter that classifying FMD participants is challenging even in the clinic. On the other hand, using the SD improves the detection of the HC group from MP group. It can be seen that the SD's UB makes it challenging to distinguish between ND and MCI.

In the two-way classification, the experiment is examined from two aspects: differentiating people with MP from each other (ND vs. MCI, ND vs. FMD, MCI vs. FMD) and distinguishing each group with MP from HC (ND vs. HC, MCI vs. HC, FMD vs. HC).

Table 5.11 Classification results of four-way classification (ND vs. MCI vs. FMD vs. HC) using two factors for SD a= 3 as the default value and a = 1.5 as the factor with the highest performance, and for IQR b= 1.5 as the default value and b=2 as the factor with the highest performance. These approaches are tested on 70 thresholds or features using linear SVM.

Factor	Approach	Accuracy	Precision	Recall	F-measure
a,b =1.5	SD (Min-UB)	44%	35%	43%	38%
	IQR	**49%**	**53%**	**49%**	**49%**
a=3, b=2	SD (Min-UB)	42%	32%	41%	35%
	IQR	**49%**	**54%**	**49%**	**49%**

The first aspect that classifies people with memory-related problems from each other is investigated, and the obtained results are shown in Table 5.12. As mentioned above, IQR gives better results in classifying MP classes. The highest performance between the three classification tasks is achieved with 72% to classify ND and MCI from each other. Using the SD causes an incorrect classification for all of the FMD group regardless of the factor value, as presented in Figure 5.15.

The confusion matrix results for the four-way and two-way problems are observed with the plots of the EAR for each participant based on their class. They show that when there are many fluctuations above the line where most of the EAR values lay, the participants tend to be classified as ND. However, when these fluctuations lay below the line, the participants tend to be classified as MCI. For instance, an FMD participant (P40) is classified as ND. Participants P01 and P10 are classified as MCI regardless of the method, SD or IQR, when the four-way classification is done.

In the two-way problem (ND vs. FMD), FMD participants (P01, P07, P09, and P40) are consistently detected as ND regardless of the approach used and the factor value. The rest of the FMD participants (P06, P10, P17 and P39) are classified as ND when the SD approach is used. The rest of the FMD participants (P06, P10, P17 and P39) are classified as ND only when the SD approach is used. This is because the SD cut the upper part of the EAR values for these classified participants as ND and the ND participants, making the EAR characteristics for both classes close to each other. When the two-way classification MCI vs. FMD is analysed, FMD participants P01, P06, P09, P10, and P40 are classified as MCI using either the SD or the IQR approaches. Other participants of FMD (P07, P17, and P39) are detected as MCI using the SD approach, which results from cutting the upper part of both participants' MCI and FMD, making their EAR behaviour look similar to each other.

The second aspect of the two-way classification is testing the system performance in classifying every class with MP from the HC class (ND vs. HC, MCI vs. HC, and FMD vs. HC). Table 5.13 shows the results of the system performance in each classification task for

Table 5.12 Classification results of the two-way classification (ND vs. MCI, ND vs. MCI, and MCI vs. FMD) using two factors for SD, a= 3 as the default value, and a = 1.5 as the factor with highest performance, and for IQR, b= 1.5 as the default value, and as b=2 the factor with highest performance. These approaches are tested on 70 thresholds or features using linear SVM (Fact.= factor; Appr.= Approach).

Classes	Approach	Accuracy	Precision	Recall	F-measure
ND vs. MCI	SD (Min-UB)	67%	68%	67%	66%
	IQR	**72%**	**72%**	**72%**	**71%**
ND vs. FMD	SD (Min-UB)	56%	29%	50%	37%
	IQR	**69%**	**68%**	**66%**	**66%**
MCI vs. FMD	SD (Min-UB)	54%	28%	50%	36%
	IQR	**67%**	**70%**	**64%**	**63%**

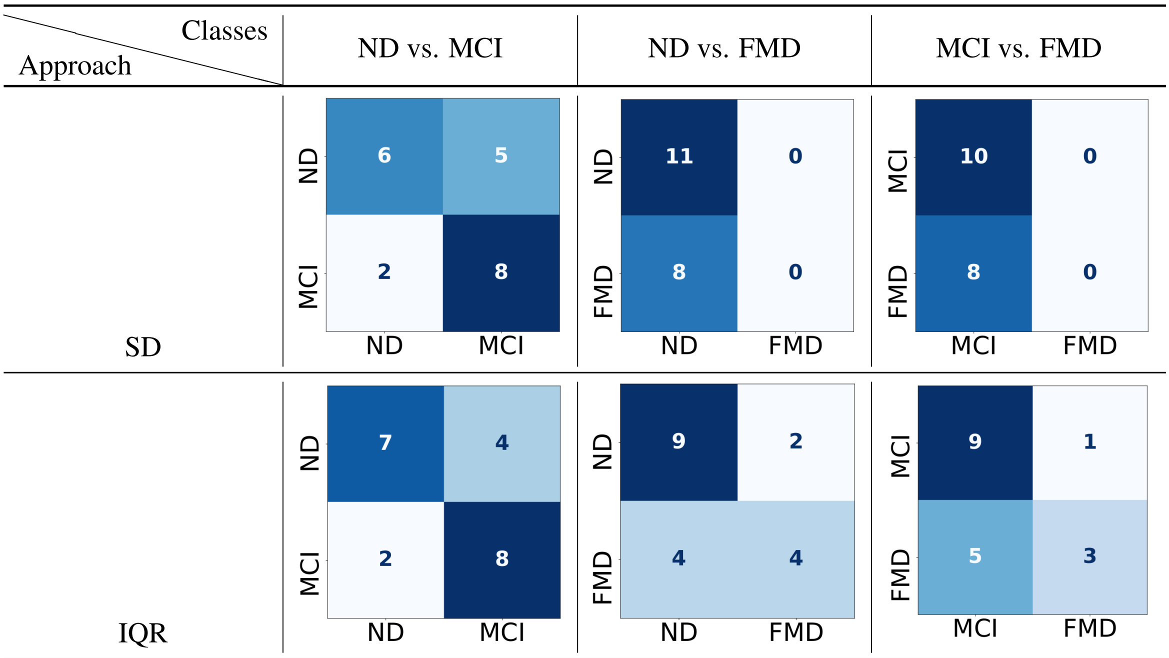


Fig. 5.15 The confusion matrices for the two-way classification (ND vs. MCI, ND vs. MCI, and MCI vs. FMD) for both the SD and the IQR , using 70 features or thresholds (In confusion matrix, rows: true labels and columns: classified labels).

both approaches and different factor numbers. It can be seen that using the SD approach gives a much better performance than the IQR approach. In ND vs. HC, increasing the threshold range gives better classification performance from factor a = 1.5 with an accuracy of 69% to a = 3 with 78%. In contrast, reducing the threshold range from a = 3 to a = 1.5 enhances the performance from 78% to 89%, respectively. In FMD vs. HC, the SD approach has the same performance when the factor changes. It can be seen in Figure 5.16 that the predictions of these classes for both approaches explain which class is classified much better. It is observed that using the SD with a high factor (a=3) increases the detection of ND participants from the HC group.

In contrast, the HC participants could be detected better using IQR regardless of the factor value. Although MCI and HC are detected significantly using the SD with a low factor value (a = 1.5), only two participants are classified incorrectly, as seen in the confusion matrix. Interestingly, it can be seen that none of the MCI participants are ever confused with HC participants when the IQR approach is used.

Regarding the FMD vs. HC classification, the HC participants are much better detected from the FMD group when the SD factor a = 1.5 or 3, and the IQR factor b = 2. The best factor for detecting FMD from HC is achieved using the high value of the SD factor (a = 3), with only two FMD participants confused with the HC group. Taking all these findings into account, SD with a high factor value could help the model in the training phase to learn by capturing a pattern that could distinguish between these classes with MP from the HC group. It can be seen that using a low value for the factor results in a model that confuses the two classes and cannot classify them from each other. In contrast, the detection of HC from other classes such as ND and FMD is much better when the IQR is used. It captures a smaller range than the SD, which leads to finding features from the model that identifies the HC from other classes.

Table 5.13 Classification results of the two-way classification (ND vs. HC, MCI vs. HC, and FMD vs. HC) using two factors for SD, a= 3 as the default value, and a = 1.5 as the factor with the highest performance, and for IQR, b= 1.5 as the default value, and b=2 as the factor with the highest performance. These approaches are tested on 70 thresholds or features using linear SVM (Fact.: factor; Appr.: approach; P: precision; R: recall; F1: f-measure).

		ND vs. HC				MCI vs. HC				FMD vs. HC			
Factor	**Approach**	**Accuracy**	**P**	**R**	**F1**	**Accuracy**	**P**	**R**	**F1**	**Accuracy**	**P**	**R**	**F1**
a,b=1.5	SD (Min-UB)	**69%**	**70%**	**69%**	**69%**	**89%**	**89%**	**89%**	**89%**	**77%**	**78%**	**76%**	**76%**
	IQR	69%	67%	66%	65%	72%	83%	72%	71%	67%	65%	65%	65%
a=3,b=2	SD (Min-UB)	**78%**	**81%**	**79%**	**79%**	**78%**	**80%**	**79%**	**79%**	**77%**	**76%**	**76%**	**76%**
	IQR	72%	71%	71%	70%	72%	83%	72%	71%	67%	65%	64%	64%

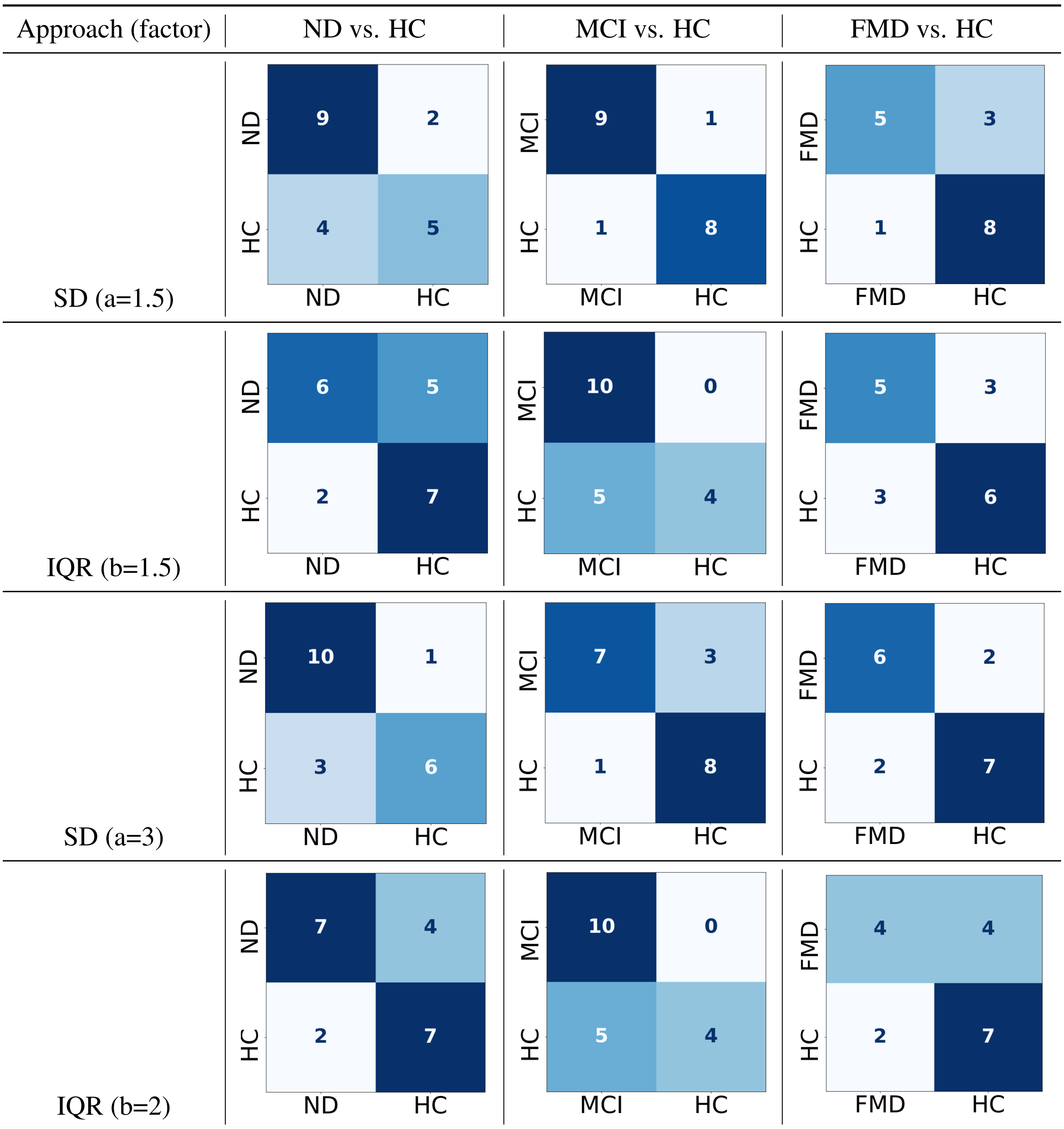


Fig. 5.16 The confusion matrices for the two-way classification (ND vs. HC, MCI vs. HC, and FMD vs. HC) for both the SD and IQR with their default factors and the ones with highest performance, using 70 features or thresholds (In confusion matrix, rows: true labels and columns: classified labels).

When the third quartile of the IQR approach is used to determine the UB, several classification tasks show better performance than when the 3rd SD is used. However, there are also several classification tasks that show better performance when the 3rd SD is used than the IQR approach. Table 5.14 summarises the different classification tasks and which UB determining approach achieves the best performance with the significance test p-value. The table shows that the IQR approach achieves a better performance on most classification tasks that detect people with MP (ND/MCI/FMD) and the four-way classification. The difference between the SD and IQR results is considered to be statistically significant. However, the SD approach shows a higher performance when any class of people with MP is classified from HC. MCI can be distinguished from HC better than the ND and FMD groups.

Table 5.14 All the classification tasks and the approach to UB determination that achieved the highest performance in the t-test to show the significant difference between the highest obtained results using the IQR and the SD (*: statistically significant; **: extremely statistically significant).

Classification Task	**Best IQR or SD?**	**P-value**
ND vs. MCI vs. FMD vs. HC	IQR	0.04*
ND vs. MCI vs. FMD	IQR	0.02*
ND vs. MCI	IQR	0.0003**
ND vs. FMD	IQR	0.02*
MCI vs. FMD	IQR	0.03*
ND vs. HC	SD	0.002*
MCI vs. HC	SD	0.01*
FMD vs. HC	SD	0.0004**
MP vs. HC	SD	0.0001**
D vs. Non-D	IQR	0.0001**

To conclude, when the classification tasks are between the MP (ND, MCI, and FMD), the IQR captures a good EBR behaviour that improves the classification task. On the other hand, when the classification tasks are carried out between any health condition with a memory problem, the SD performs better at capturing behaviour that can distinguish between healthy people and any health condition. The reasons for this can be related to the data variations in the recording environments and the previously mentioned challenges that lead to a different EAR range behaviour and the data size. Therefore, further work needs to be done on a larger dataset.

5.7 Discussion and Conclusion

In this study, experiments were conducted to investigate the feasibility of using the MTs approach on the extended data given the variations in the recording environments and of classifying people with MP (ND, MCI and FMD) and HC as a four-way, three-way and a combination of two-way problems. When the MTs-OAP approach was applied to the combined dataset IVA_{52}, it showed an issue in the prediction of MCI participants, who were all classified as ND. Even applying the approach of the two-way classification to distinguish HC from any memory-related problem classes showed that all the new data was misclassified as HC. The reason behind this was the variation in the recording environments and the different devices used - laptops and smartphones. In addition, the mean and the SD of the new data participants were lower than in the IVA_{18} data.

As a consequence of these issues, the MTs-OAP was improved by making the MTs calculation PD (i.e., the thresholds for each participant were calculated based on their minimum and maximum values). However, using the maximum did not show good results because it took the very extreme values into account and resulted in many zero or close-to-zero EBR values. Therefore, the SD approach to determine the UB and LB was utilised. This approach showed better results than the min-max method, but the LB caused an issue by including more unneeded thresholds that resulted in many zero values. Thus, the LB was considered the minimum value of the EAR. Then, the performance was measured based on the new setup for the three-way and two-way classification tasks. The resulting performance showed improvements only in classifying HC from MP, which could have been due to the SD sensitivity of extreme values and thus affected the UB calculation.

To overcome the problem of the SD, the IQR approach was investigated to determine the UB because it is very robust in handling data with extreme values. The IQR improved only the performance of the three-way classification and the prediction of the D group from the Non-D group. Moreover, the IQR achieved better results even in the four-way (e.g., ND vs. MCI vs. FMD vs. HC) and two-way classification tasks that predicted ND, MCI and FMD from each other, whereas SD did not perform well, and the results were reflected in the confusion matrix when the classification task involved FMD. The confusion matrix showed that all FMD participants were misclassified, indicating the difficulty in distinguishing FMD from either ND or MCI.

On the other hand, when the HC was classified from either ND, MCI, or FMD, the SD gave relatively better results than IQR. This shows that IQR was better for classification tasks related only to MP. However, SD was better to use on a classification that involved classifying HC from any memory-related problems. That is to say, the SD range was larger than IQR, and this could help to capture better patterns to be able to predict HC from other classes, whereas the IQR range could help to capture a pattern between the memory-related problems classes (e.g., ND, MCI and FMD).

Classifying these three classes, ND, MCI, and FMD, from each other in three-way and two-way classification tasks was very challenging because previous studies had only classified dementia regardless of the type from healthy controls. This study showed promising results using only one feature, EBR. The results could be improved by investigating the performance with other visual features, and this will be described in the next chapter.

Chapter 6

Head Movement Based Detection of Cognitive Impairment

> *"You are not alone. You are seen.*
> *I am with you. You are not alone."*
> *— Shonda Rhimes*

6.1 Introduction

The previous chapter focused on eye blink rate (EBR) to differentiate neurodegenerative disorder (ND), mild cognitive impairment (MCI), functional memory disorder (FMD) and healthy controls (HC). This chapter investigates an additional visual cue, head movement, in detecting CI and related health conditions on the IVA_{18} dataset. Then, the system is evaluated on a larger dataset, IVA_{52}, which is a combination of IVA_{18} and IVA_{34} datasets, as described in Chapter 3.5. A growing body of literature recognises the importance of head movement in human communication (Boholm and Allwood, 2010; Hadar et al., 1984; Maynard, 1987; Mehrabian, 2017). It is the easiest non-verbal cue to understand and conveys rich information. Previous work has investigated the role of head movement during conversation (Maynard, 1987) and communicative feedback (Boholm and Allwood, 2010) and found that people often use nods (up then down, or down then up), tilt (leaning the head to one side), and head turn (left then right, or right then left) for emphasising a word and for yielding or asking for a turn in the conversation (Hadar et al., 1984), indicating an agreement or disagreement (Boholm and Allwood, 2010), expressing a positive or negative attitude, showing how anxious a person is (Mehrabian, 2017), helping in assessing pain in patients (Werner et al., 2018) and evaluating mental state (Larner, 2005, 2012).

Given its importance in social communication, head movement analysis is increasingly important in applied computer vision. Recently, a number of studies developed automatic facial behaviour analysis tools that involve head movement (Al-Rahayfeh and Faezipour, 2013b; Baltrusaitis et al., 2018) to facilitate the understanding of a person's mental status, such as slower head movements, less change of head position and longer duration of looking down and to the right in depression (Alghowinem et al., 2013b), longer duration of looking down in post-traumatic stress disorder (Stratou et al., 2013), slower head movements as a suicidal cue (Laksana et al., 2017), and reduced head movement in schizophrenia (Jiang et al., 2022). Other fields have benefited from the automatic analysis of facial behaviour, including education (McDaniel et al., 2007) and the automotive industries (Busso and Jain, 2012).

So far, head movement analysis focusing on CI has received scant attention in the medical research literature. A number of cross-sectional studies have suggested an association between CI and a particular head movement, which is the head turn (Fukui et al., 2011; Larner, 2005, 2012, 2014b). Larner (2005, 2012, 2014b) and Soysal et al. (2017) found that the number of head turns and the presence of a partner or other people with the patient can be considered clinical cues for CI and dementia. In addition, Fukui et al. (2011) showed that the head turn cue indicates CI regardless of a partner's presence. These studies showed that the head turn cue is a strong cue on its own. This difference in the findings may be due to the different ways in which the researchers divided the participants into groups. Fukui et al. (2011) divided the participants into two groups: AD-related, which included Alzheimer's disease (AD) and MCI, and AD-nonrelated, representing dementia with Lewy bodies (DLB), progressive supranuclear palsy (PSP), and vascular dementia (VaD), whereas Soysal et al. (2017) considered both AD-related and AD-nonrelated as one group.

Very little attention has been paid to investigating the automatic detection of CI using head movement. This thesis, therefore, aims to explore the relationship between CI and head movement cues and assess the significance of this behaviour in identifying CI and related health conditions. This research explores, for the first time, the automatic detection of CI based on the head turn cue and more general head movement analysis and explains how these cues can be effective in measuring the severity of CI. In addition, this is the first study to use data recorded in the wild involving people with health conditions.

This chapter first analyses the head movement data in Section 6.2. Then, Section 6.3 covers the feature extraction of the head movement and describes the feature fusion of head movement with the EBR feature, previously explained in Chapter 4. Section 6.4 provides a brief overview of the classifiers and evaluation metrics used in this research. The research findings are covered in Sections 6.5. Section 6.6 evaluates the system on the IVA_{52} dataset. The findings are then discussed in 6.7. Section 6.8 finally presents a summary of the chapter.

6.2 Analysing Head Movement

In this chapter, the IVA_{18} and IVA_{52} datasets are used, which are the same as were used in Chapter 4. After the data has been pre-processed as described in Section 4.2.1, the OpenFace toolkit for facial landmarks tracking, presented in the previous chapter, was used to track the facial landmarks and estimate the head movement for each frame. This resulted in three Euler angles: pitch, yaw and roll (see Figure 6.1). This section provides an analysis of the participants' behaviour in terms of head movement for the IVA_{18} and IVA_{34} datasets, which can be influenced by their health condition, the presence of a partner, recording environments and devices used. Then, the analysis of the IVA_{34} dataset is compared with the analysis of the IVA_{18} dataset. In addition, this section describes the challenges observed in the head movement data.

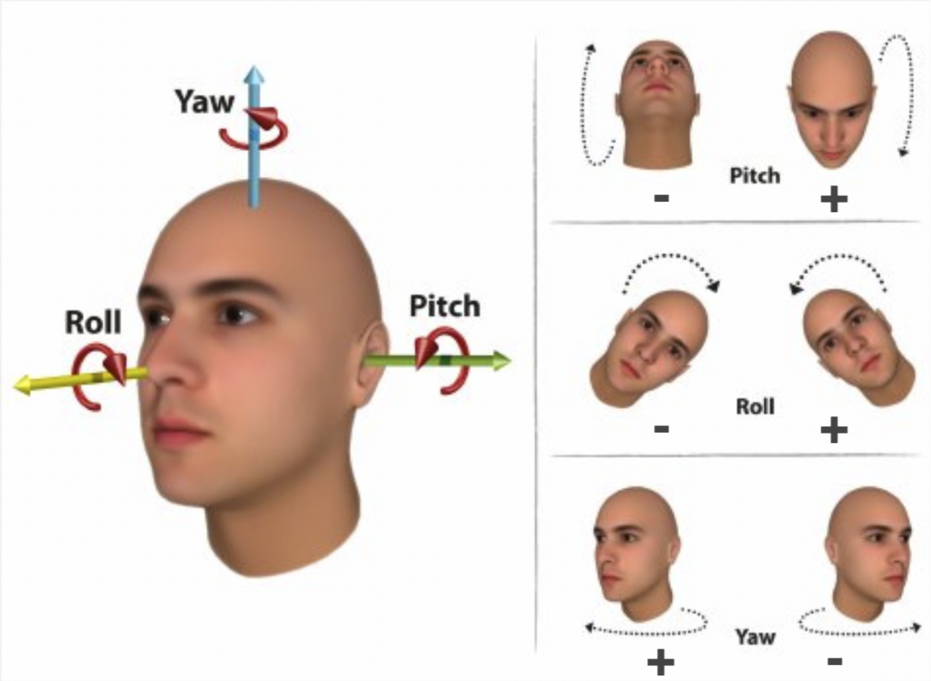


Fig. 6.1 Head movement showing the three degrees of freedom: pitch, yaw, and roll, adopted from Arcoverde Neto et al. (2014).

6.2.1 Analysing Participants' Behaviour for the IVA_{18} Dataset

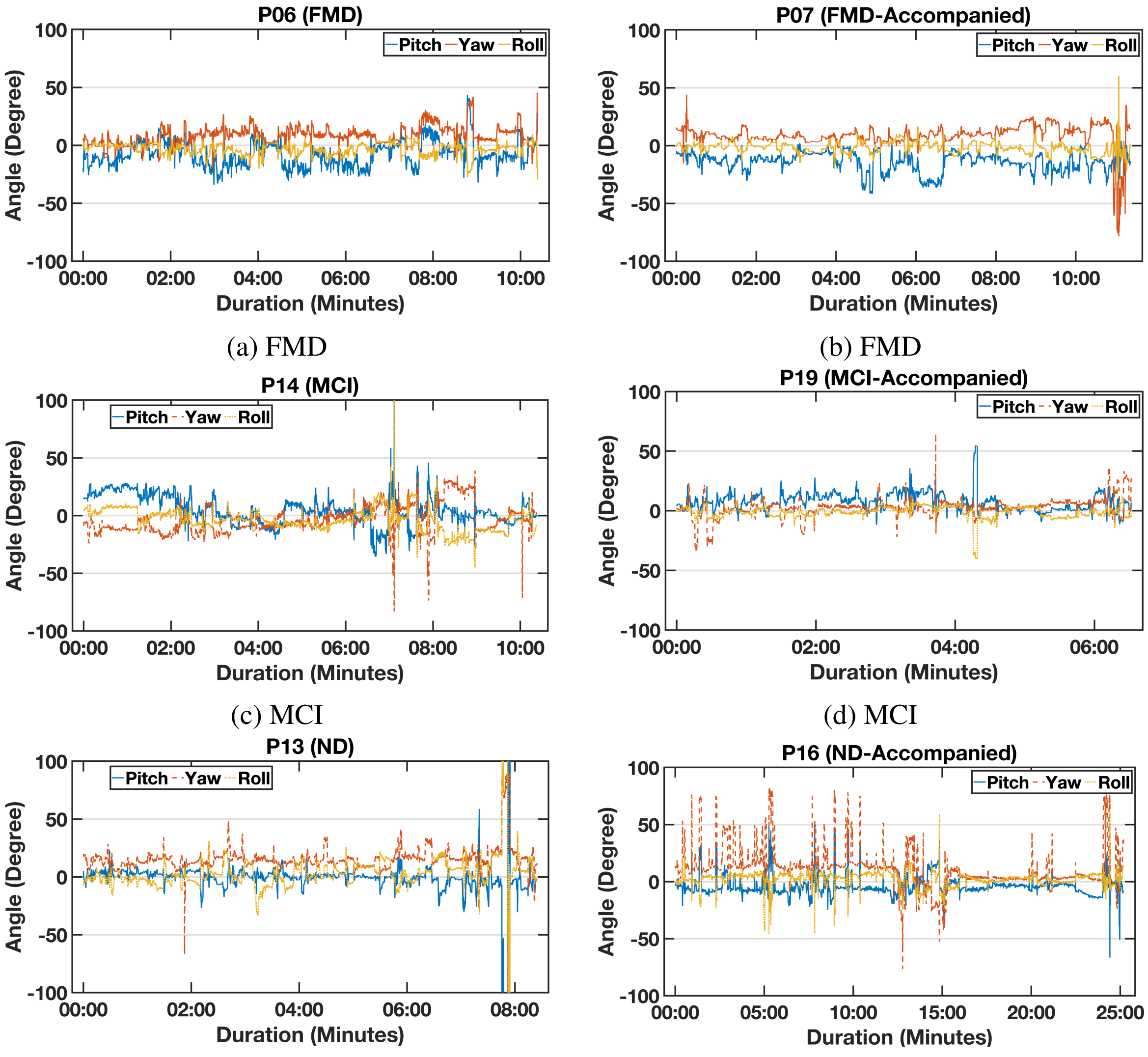
Figure 6.2 presents the head movement data (i.e., Euler Angles: pitch, yaw, and roll) of six participants, two participants in each group. The three calculated angles of the head orientation are analysed for each group. Appendix B.2.3 shows figures for all of the participants in each group. The x-axis and y-axis represent the session duration and the angle value, respectively.

The first row of Figure 6.2 shows the general pattern of participants with FMD. It can be seen that the pitch angle, which declines significantly for several participants, could indicate that they move their heads up trying to recall or think of the answer. The middle row of the figure shows that the head movement of participants with MCI differs from those with FMD in several respects. Participants with MCI show an increase in head turns and head movement than participants with FMD. Moreover, high variations often appear during most sessions, and the head faces downwards more than upwards. The most obvious difference is in the pitch angle, which increases significantly (head faces downwards) for some participants, which might indicate a state of thinking or difficulty recalling a memory. What stands out in the third row of the figure, which is for participants with ND, is the high rate of variations of the head movement from the other two groups throughout the whole session. In particular, the head turn rate (HTR) increases more with higher values in the direction of where the accompanying person or the doctor sits, and the direction of pitch value does not seem specific to any one direction.

Interestingly, the presence of a partner factor plays a crucial role in increasing the participants' head turns and movement, as shown in the right-hand column of the figure for participants who came with an accompanying person. The impact of this factor can be seen more in participants with ND (P11, P16, P22 and P23), who have higher HTR than the other two groups, MCI and FMD. In addition, the MCI group has higher HTR than the FMD group, which can be seen in participants with MCI (P15 and P19) and FMD (P07). These observations indicates that increases in the HTR may relate to the severity of the CI, which means that when an MCI participant has a very high HTR, it could signify a progression state from MCI to AD (Durães et al., 2018; Larner, 2018).

6.2.2 Analysing Participants' Behaviour for the IVA_{34} Dataset

Figure 6.3 shows the estimated head movement in the three angles of pitch, yaw and roll of eight participants, two in each group (ND, MCI, FMD, HC). These three angles are analysed for each group. Appendix B.2.3 shows figures for all of the participants in each group. The x-axis and y-axis represent the video frames and the angle value, respectively. The first row of Figure 6.3 presents the general pattern of HC participants. Most HC participants tend to move their heads up during the session, indicating a memory recall or thinking of an answer. The second row in the figure presents the head movement for participants with FMD, which show different behaviour from each other. It is, however, difficult to make an analysis from the data for only two participants. The third row in the figure shows the head movement



(a) FMD

(b) FMD

(c) MCI

(d) MCI

(e) ND

(f) ND

Fig. 6.2 The calculated pitch, yaw and roll values for participants with FMD, MCI and ND.

of participants with MCI. It can be seen that these participants, especially participant P84, show a significant increase in head turns and movement than participants with FMD and HC. The reason for this is the partner presence factor because that participant came with a partner. In addition, the head turns can be seen to the left side where the partner sat. The fourth row, which is for participants with ND, shows a higher variation in head turns than both of the other groups because all of them came with partners, which is the same reason for the MCI participant. The HTR, particularly, increases more in the direction of where the accompanying person sits, and there is no particular direction of the pitch value.

Taking all the analysis mentioned above, the partner presence plays a significant role in increasing the head turn and movement of a participant, which is consistent with the previous analysis for the IVA_{18} dataset described in Chapter 6. In addition, the HC participant in the IVA_{34} dataset shows similar behaviour to the FMD group in terms of pitch direction. The IVA_{34} dataset includes many challenges, as explained in Chapter 3, that result in extremely high values as outliers in the head movement calculation, as shown in Figure **??**. This problem is resolved by detecting and removing any value that is $\pm 90^\circ$ and then using a linear interpolation approach to fill the gaps where the outliers are.

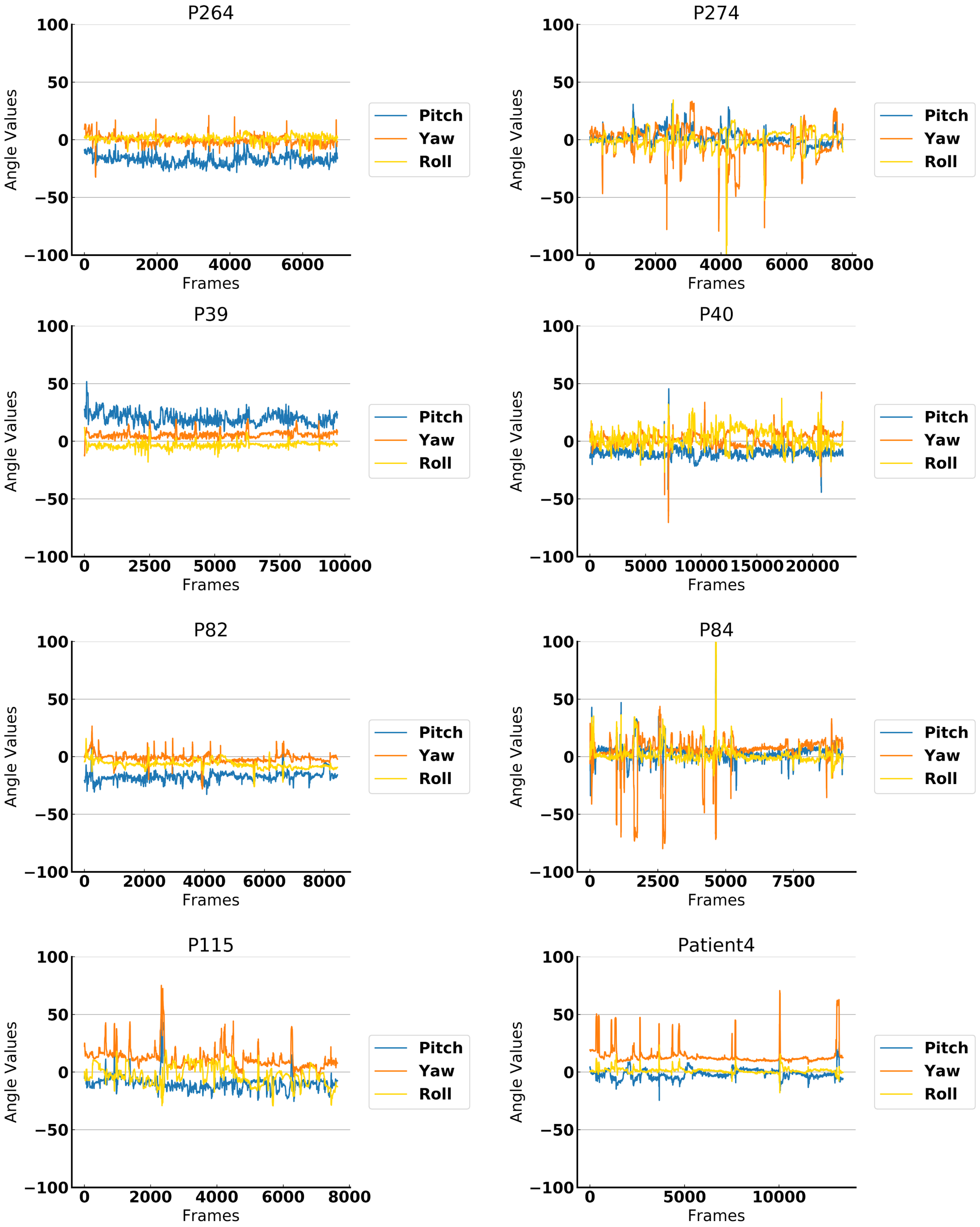


Fig. 6.3 An example of the estimated three angle values for different participants based on their diagnostic class (first row: HC participants, second row: FMD participants, third row: MCI participants and fourth row: ND participants).

6.2.3 Challenges Observed in the Head Movement Data

Head movement estimation has been studied by many researchers using different approaches (Baltrusaitis et al., 2018; He et al., 2022; Hsu et al., 2018; Khan et al., 2019; Kuhnke and Ostermann, 2019). Redondo-Cabrera et al. (2016) found several challenges that can lead to false positive values in the calculation of the head movement. The types of false values are categorised as: (1) *correct* if the error is $< 15°$, (2) *opposite* if the error is $> 165°$, (3) *nearby* if the error is $[15°, 30°]$; and (4) *other* for all other situations. In the IVA_{18} data, two types of these false values are observed, which are *other* and *opposite*. In this section, the terms *'opposite'* and *'other'* are referred to as *'outlier'* and *'false head turn'*, respectively. The following gives an example of the observation and analysis conducted on one participant.

False Head Turn

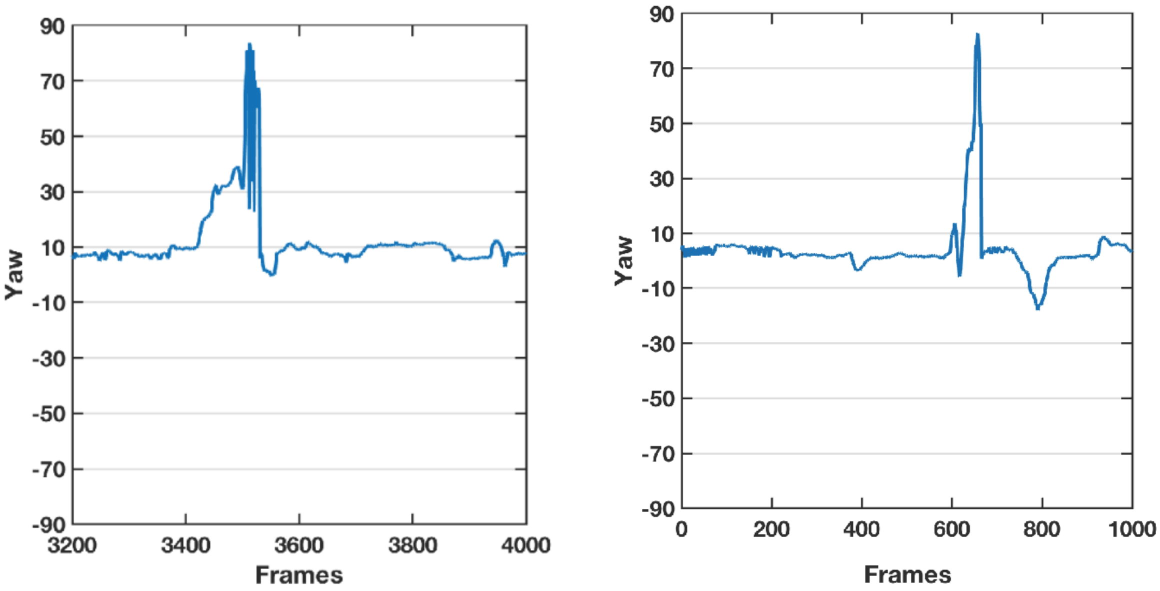
The video of this participant is observed frame-by-frame in terms of whether the high values indicated genuine HTs or not. Figure 6.4 presents the yaw angle over time for participant P11. Both subfigures show a head turn. The observed high value (in the left-hand figure) is a genuine turn that involves fast head movement to the right side. However, the high values of the estimated angle (in the right-hand figure) are a false head turn due to the effect of the face being occluded by a piece of paper. This may show how much occlusion issues have affected the calculation of the head movement depending on the amount of the occlusion and how the efficiency of facial landmarks is detected. When the occlusion is reasonably small, it will not cause a problem. However, the occlusion values of the head pose and movement angles would not be reliable.

This issue can be seen in the IVA_{18} data due to the nature of the recording settings and cannot be resolved unless each video is manually inspected frame-by-frame to check whether each head movement is true or false. However, the proportion of this kind of false head movement value was found to be very small after the videos were manually observed. It is therefore assumed that their impact is minimal.

Outliers

Another issue is the high values that exceed the normal range for head movements, which is $\pm 90°$. Unlike the previous issue of determining whether a head movement is true or not, the latter issue can be resolved, as will be described as follows.

Three participants in the IVA_{18} data have a problem related to the estimated head pose angles, resulting in extremely high values, which can be regarded as outliers. These participants were one from the MCI group and two from the ND group. Outliers are unreliable values that



(a) True head turn

(b) False head turn

Fig. 6.4 A zoomed part of the detected frames of participant P11 in the IVA_{18} dataset to show the detected yaw value when there is a true head turn (a) and a false head turn (b).

may appear due to the calculation when a participant covers his/her face with his/her hands, as shown in Figure B.7a for participant P11. Figures B.6c, B.7a, and B.7b in the Appendix show that these values are above 120°, which cannot be true given it is impossible to move the head more than about 90° (Gourier et al., 2004; Gross et al., 2008). A test is, therefore, carried out to determine the limit of the head movement, which the Openface can capture by recording short videos of myself during the day and at night to measure how moving the head in different lighting conditions may affect the calculation of head movement (see Appendix B). The experiment shows that fast head movement and head movement in poor lighting conditions can result in outlier values.

The data outliers that must be addressed are caused by an extreme head movement that is difficult to deal with using OpenFace. They are detected and removed when the value of the pitch, yaw or roll angle is ±90°, following previous work (Cao et al., 2018; Huang et al., 2008; Koestinger et al., 2011; Liu et al., 2016). Then, a linear interpolation process is used to fill in the gaps where outliers occur, as shown in Figure 6.5.

6.3 Feature Extraction

This section describes the extraction of four types of visual features: head turn rate (HTR), head turn statistical features (HTSF), which is related to the HTR feature, head movement

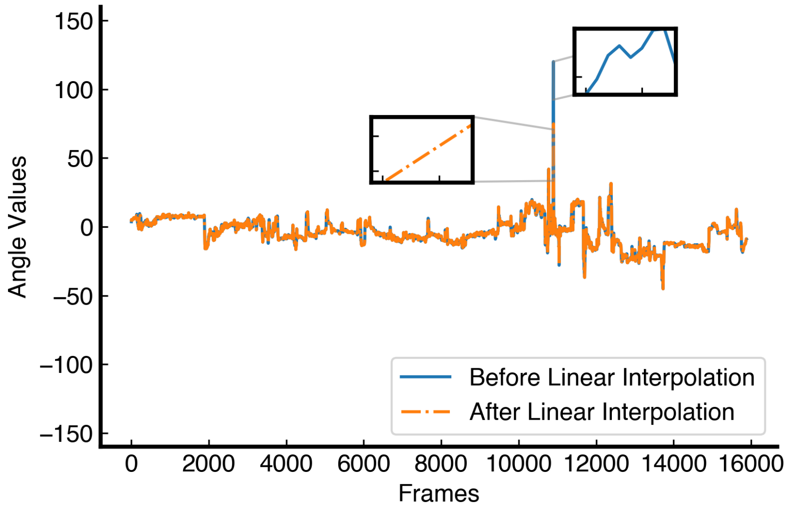
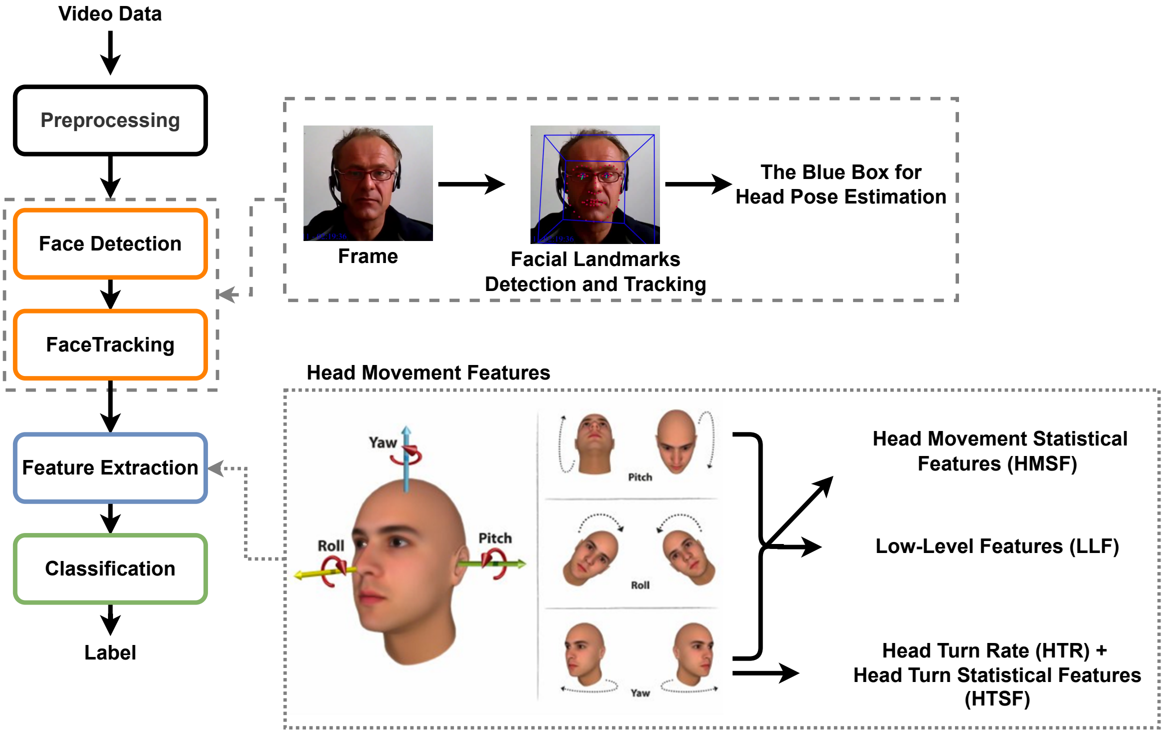


Fig. 6.5 An example of linear interpolation for a roll angle for a participant with MCI

statistical features (HMSF) and low-level features (LLFs) (see Figure 6.6). Each feature type is explained in detail in the following sections.



Video Data
Preprocessing
Face Detection
FaceTracking
Feature Extraction
Classification
Label
Frame
Facial Landmarks Detection and Tracking
The Blue Box for Head Pose Estimation
Head Movement Features
Yaw
Roll
Pitch
Pitch
Roll
Yaw
Head Movement Statistical Features (HMSF)
Low-Level Features (LLF)
Head Turn Rate (HTR) + Head Turn Statistical Features (HTSF)

Fig. 6.6 Pipeline of the extraction of the head movement features for each participant.

6.3.1 Extraction of Head Turn Rate Feature (HTR) and Head Turn Statistical Features (HTSF)

As stated above, many medical researchers have utilised the head turn cue to investigate its association with a decline in the CI (Fukui et al., 2011; Larner, 2012, 2014b; Soysal et al., 2017). This cue is adopted in this work to explore an automatic approach for its extraction and then CI detection. This exploratory study can help to obtain further in-depth information on the relationship between the head turn cue and CI and the possible factors that may have an effect on this cue. As explained above, OpenFace was used to extract the yaw angle for each participant. With this angle, the head turn to the left or right side is detected, and additional features are related to each detected head turn.

The SciPy [1] package is used to detect signal peaks and corresponding parameters values, following similar studies (Wei et al., 2021; Zheng et al., 2020). The algorithm of peak detection takes the yaw angle as input and searches for the local maxima by a simple comparison of intensity. The following properties are used to select a subset of the peaks: peak prominence, peak height and peak width. Peak prominence measures a peak's significance based on its intrinsic height and relative location from surrounding peaks, as shown in Figure 6.7. The definition of height and width can be found in the same figure. Throughout the yaw angle example, the same parameter is used for peak detection with prominence = $\pm 45^\circ$, which means that any value below or above 45° is considered a peak. Each detected peak counts as a head turn. The HTR is calculated by dividing the number of head turns by the number of frames. More derivative features related to the peak detection are extracted, such as the actual values of peak prominence, peak heights, peak widths and the distances between peaks, because they may be useful for identifying a pattern of head turns to differentiate the groups from one another.

Then, statistical features, the mean, Standard Deviation (SD) and Variance (Var), are calculated for each derived feature. These statistical features are used as complementary cues with HTR and are referred to as HTSF (see Table 6.1). Previous work has used these kinds of statistical features (Lee et al., 2020; Valstar et al., 2014, 2013).

6.3.2 Extraction of the Head Movement Statistical Features (HMSF)

Many researchers have used head movement for depression detection (Alghowinem et al., 2013b; Ringeval et al., 2018, 2019). However, no research has been conducted to investigate the association between head movement and CI. This research investigates the head movement as a cue to evaluate its effectiveness in automatically identifying CI. Following similar related

[1] https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find peaks.html

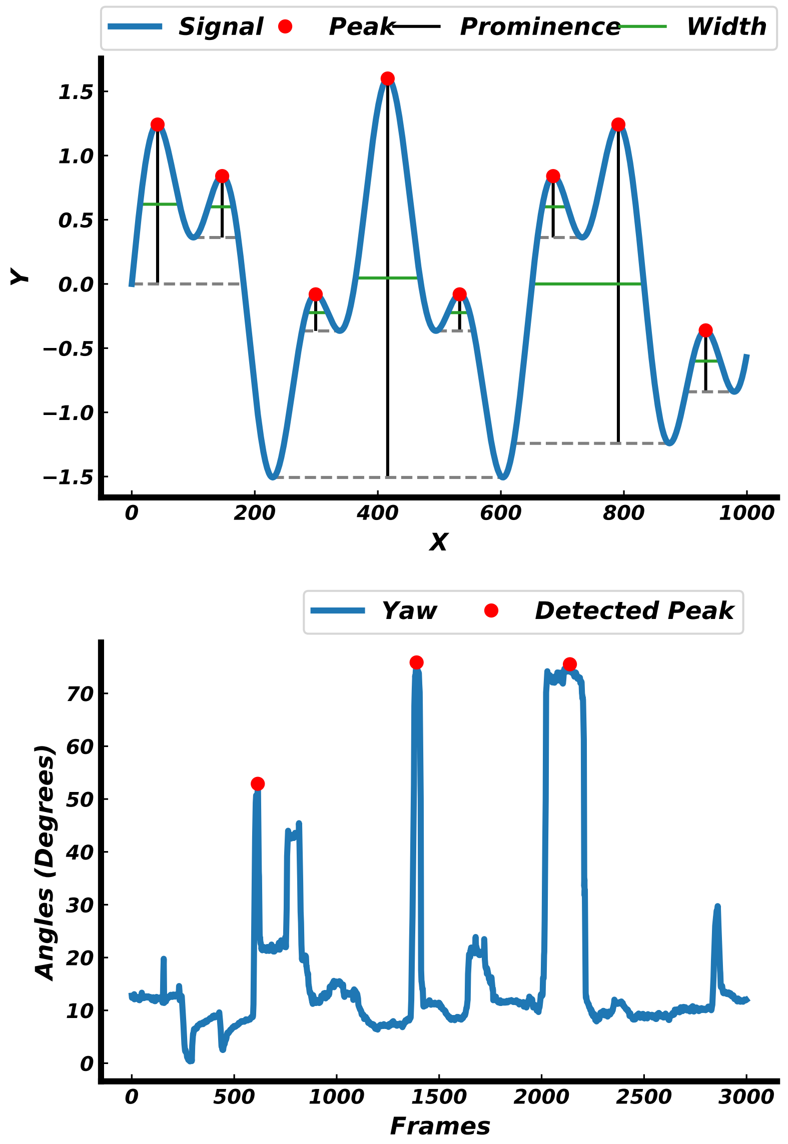


Fig. 6.7 Examples of peak detection in the SciPy python package. (top) a diagram depicting the peak prominence definition., which is defined as the vertical line between the peak and the lowest contour line (the grey line), and (bottom) peak detection from the yaw angle of one random participant.

work (Alghowinem et al., 2013b), this section explores the use of all the angles of the three head movements (i.e., pitch, yaw, and roll), including the velocity and acceleration for each angle. The velocity is defined as angular velocity (ω) that measures how fast the object rotates over time (Kimball, 1917). The angular velocity can be calculated as angle per unit time, e.g., radiance per second (θ/t). The angular acceleration (α) is the change in the angular velocity per unit time. Based on this definition, the equation is $\alpha = \Delta\omega/\Delta t$. Then,

Table 6.1 Description of each head movement feature

Feature type	**Feature description**	**Feature Abbrev. (dimension)**
Head turn (yaw)	head turn rate	HTR (1 X 1)
	head turn statistical features: head turn prominence-mean, prominence-SD, prominence-Var, height-mean, height-SD, height-Var, width-mean, width-SD, width-Var, distance-mean, distance-SD, and distance-Var.	HTSF (1 X 12)
Head movement statistical features (pitch, yaw, and roll)	minimum, maximum, mean, range, SD, and Var are calculated for each angle, velocity and acceleration vector.	HMSF (9 X 6)
Low-level features (frame-by-frame)	the three angles, velocity, and acceleration	LLF (9 X 12)

the following statistical features for each angle, velocity, and acceleration are calculated: mean, SD, variance, range, maximum and minimum, as shown in Table 6.1. This results in 54 features (6 x 9), referred to as HMST.

6.3.3 Extraction of the Low-Level Features (LLFs)

Features previously explained are calculated per video. However, another approach to extracting features is based on a frame-by-frame analysis, referred to as low-level features (LLFs). One of the problems with using frame-by-frame features is the variable length of videos. To resolve this issue, a Gaussian mixture model (GMM) is used, dimensionality reduction technique (Bishop and Nasrabadi, 2006; Murphy, 2012). This technique takes features with high dimensionality and reduces them to a lower dimension. Alghowinem et al. (2013b) and Alghowinem et al. (2015) have used a similar approach.

A GMM approach calculates the probability of each datapoint being generated from a mixture of a finite number of Gaussian distributions. Each Gaussian is identified by $k \in 1,...,K$, where K is the number of clusters of the LLF. In the mixture, each Gaussian k consists of three parameters: a mean (μ), a covariance ($\sum$) and a mixing probability (π). The μ represents its centre, the $\sum$ represents its width, and the π defines the size of the Gaussian function, whether big or small. Each Gaussian describes the data included in each cluster.

The mixing probabilities must match this condition in Eq. 6.1:

$$\sum_{k=1}^{K} \pi_k = 1 \tag{6.1}$$

The optimal parameters need to be determined to ensure that each Gaussian fits the datapoint of each cluster. To achieve this, an expectation maximization (EM) algorithm is needed to estimate unknown variables by observing and deriving them from other observed values. The EM algorithm is used to maximise the likelihood with two steps that need to be calculated: the Expectation step (E-step) and the Maximisation step (M-step). In the E-step, the posterior value of unknown data needs to be estimated.

Eq. 6.2 can be used to calculate the posterior distribution γ of every Gaussian for each datapoint z_{nk} in Eq. 6.3. This equation uses a Bayes' rule where π is a prior weight, N is the number of observations, and x_n is observation n:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sum_k) \tag{6.2}$$

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n|\mu_k, \sum_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_k, \sum_k)} \tag{6.3}$$

Then, the M-step calculates the initial parameters, which are the mean, the covariance and the weights in Eq. 6.4, 6.5 and 6.6, respectively, and then sums these parameters for each Gaussian and the marginal likelihood for maximising in Eq. 6.7, and 6.8, respectively, for each Gaussian using the following equations:

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \tag{6.4}$$

$$\sum_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k^{new})(x_n - \mu_k^{new})^T \tag{6.5}$$

$$\pi_k^{new} = \frac{N_k}{N} \tag{6.6}$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{6.7}$$

$$\ln p(X|\mu, \sum, \pi) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \sum_k) \right\} \tag{6.8}$$

$$L = \sum_{n=1}^{N} \sum_{k=1}^{K} E\left[z_{n,k}\right] (\log \pi_k + \log N(x_n|\mu_k, \sigma_k)) - \sum_{n=1}^{N} \sum_{k=1}^{K} E\left[z_{n,k}\right] \log E\left[z_{n,k}\right] \tag{6.9}$$

The data is represented as x_n with dimension $(m \times d)$; m is the number of frames, and d is the number of features given to the fitting method. The first E-step in the fitting method does the initialisation of the means and covariances to calculate *gamma*, the posterior distribution (γ). Then, the M-step takes the initial parameters and re-calculates them, corresponding to Eq. 6.4, 6.5, and 6.6 based on the new γ value and the x_n. After they have been calculated, they are used to calculate the lower bound corresponding to Eq. 6.9.

6.3.4 Normalisation

Certain pre-processing steps on the inputs of the classifier model and targets can make training the model more efficient by transforming inputs into a much better format for the model (Jayalakshmi and Santhakumaran, 2011). Prior to the training phase, the normalisation of the raw input data can have a great impact later on the training and the model performance. It can scale each input feature of the data into a particular range to reduce bias within the classifier model from one input feature to another, and speed up the training time. Several popular techniques can be used, such as Z-Score, Min-Max, Median and Sigmoid Normalisation.

In this research, normalisation is carried out for all features to scale them, which leads to Mean = 0 and SD = 1 using a Min-Max scaler (see Figure 6.8). This approach, which is a linear transformation, transfers each feature to a range between 0 and 1. This kind of normalisation can preserve the relationships in the data and only change the data scale. It does not affect the outliers.

6.3.5 Feature Fusion

Performance is measured using each feature individually and then when fused, as shown in Figure 6.8. The features are fused by concatenating them in several ways: EBR + HTR, EBR + HTR + HTSF, EBR + HMSF, and finally, the fusion of all the facial features. The fusion of features can improve the classification performance over using just one feature and can provide more helpful information for small datasets such as the one used in this study.

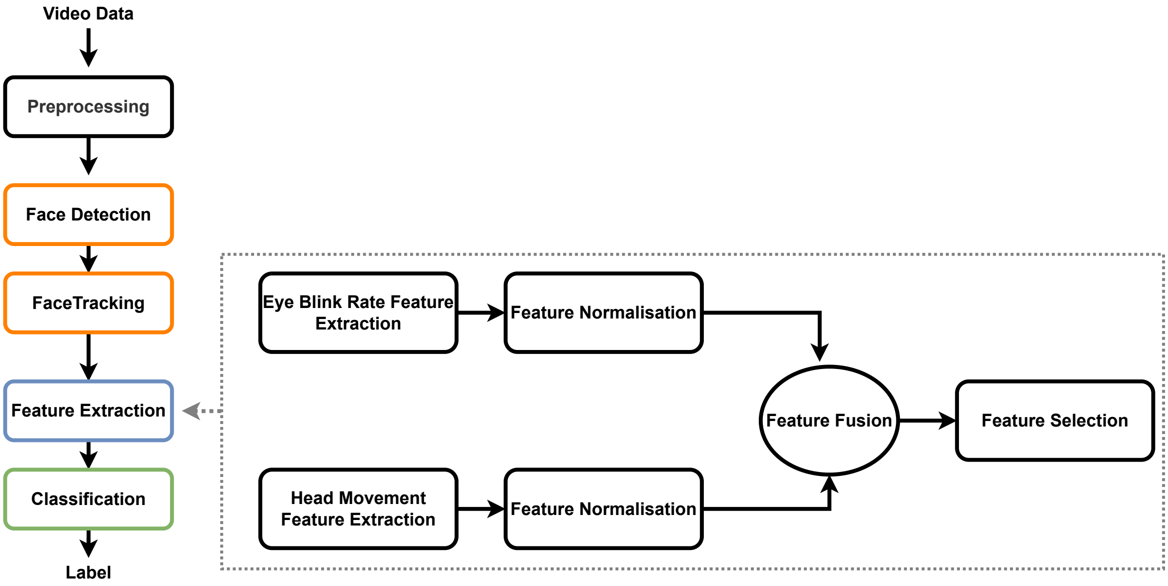


Fig. 6.8 Pipeline of the visual features normalisation, fusion and selection.

6.3.6 Feature Selection

Feature selection includes selecting the most important features in a classification problem and removing the irrelevant ones (Bishop and Nasrabadi, 2006). The effects of applying feature selection to the combined features are then evaluated. Therefore, feature selection is considered one of the feasible solutions that can help to reduce the training time and the overfitting and thus to improve the classifier's performance. A number of feature selection methods for classification purposes can be divided into three categories: filters, wrappers and embedded methods (Guyon et al., 2002; Pudjihartono et al., 2022). The main differences between these approaches are 1) if the feature selection depends on the classifier or not, 2) the metrics used for evaluation, 3) the complexity of the computation, and 4) the potential to detect the dependencies between features.

Filter-based approaches use various statistical tests (e.g., the chi-squared test, person correlation and t-test) to measure their correlation with the class and use feature ranking as an evaluation metric. Wrapper-based approaches (e.g., recursive feature elimination, KBest and randomised hill climbing) select a subset of features with the best performance for the chosen classifier and use this classifier's performance as an evaluation metric. Embedded-based approaches (e.g., random forest and lasso (L1)) integrate the feature selection process and the model construction into a single step– the classifier adjusts its parameters during the training phase by measuring the appropriate weight provided for each feature to result in the best accuracy of classification.

Pudjihartono et al. (2022) stated that choosing which feature selection method is best to use depends on the dataset used and the aim of the researcher. For instance, if the researcher's goal is to quantify the most significant features to uncover the biological mechanism behind a particular disease, then filter-based methods are the best to use because they can provide a ranked list of the most computationally efficient features. However, when the feature dimensionality in the dataset is relatively low (e.g., tens to hundreds of features), then wrapper-based methods are likely to provide the best performance, such as KBest and Recursive feature elimination cross-validation (RFECV). This research has explored t-test, KBest, and RFECV approaches for feature selection and found that the RFECV gave the highest performance. RFECV is, therefore, adopted in this research because it implicitly considers feature dependencies, which involve interaction with the classifier and redundancies during the feature selection process. In addition, it provides better performance than filter-based methods and produces the best feature set instead of a ranked list of features. RFECV is a common method that obtains the important features, ranks them, discards the irrelevant ones, and re-fits the model until a particular number of features remains. The feature selection is applied to every feature fusion mentioned previously in Section 6.3.5.

6.4 Analysis and Evaluation

Classification

As was explained in Section 4.2.7, classification involves binary and multi-class classification problems. The same classifiers used previously as described in Section 4.2.7 are used in this chapter: support vector machine learning classifier (SVM) with both linear kernels (L-SVM) and RBF (RBF-SVM), k-nearest neighbours (KNN) with uniform and distance weights, logistic regression (LR) and decision trees (DT). The hyper-parameters are optimised for each classifier using a grid search to find the best parameter. Participant-independent-stratified cross-validation (CV) is used. The best-performing parameters are estimated for each fold, and then the average is taken across all the folds. CV with 6-fold is used for each classifier.

The evaluation is then conducted using a participant-independent-stratified CV with 6-folds, as described in Section 4.2.7. In each fold, three participants are held out as a test set, and all the remaining participants are used in the training set. Each fold is split to maintain the sample distribution in each class. Confusion matrices are used to analyse the classification results. Since the data is balanced, the accuracy metric is used.

Statistical Analysis

The extracted statistical features of the head movement from the IVA_{18} data can provide valuable information about behavioural patterns. Statistical significance tests are, therefore, conducted across the ND, MCI and FMD groups. The analysis is carried out as a binary task (ND/MCI, ND/FMD, and MCI/FMD). The first step in this process is to perform a parametric test to determine whether the feature is normally distributed or not (Siebert and Siebert, 2017). Several statistical tests can be used to test the normality of the data, such as the Shapiro-Wilk test (SW), D'Agostino's K-squared test and a Quantile-Quantile Plot. The SW test is used in this research, which uses the frequency of the data to test the normality. This test is preferred for limited data size (Shapiro and Wilk, 1965). Then, a parametric test (two-tailed T-test) is used for normally distributed features, and a non-parametric test (the Wilcoxon T-test) is used for features that do not follow a normal distribution, with a significance of $P = 0.05$.

6.5 Experimental Results on the IVA_{18} Dataset

As stated earlier, the specific objective of this study is to investigate visual features to differentiate between ND, MCI and FMD. This section presents the experimental results of the 3-way and 2-way classification problems on the IVA_{18} dataset. The performance is measured using: 1) individual features (i.e., HTR, HTSF, HMSF and LLF) with statistical analysis, 2) feature fusion of the head movement and the EBR features, which was described in Chapter 4, and 3) feature selection for each level of fusion. Finally, a feature dimensionality transformation is conducted on the LLF (i.e., the angles, velocities and accelerations).

6.5.1 Effect of Head Movement Features

The performance of each feature is investigated with several machine learning classifiers. Table 6.2 presents the results of the classification accuracy for 3-way and 2-way classification problems. It is clear from the table that the best results obtained are with the HTR+HTSF, at 67% for the 3-way classification problem. Some features are better at distinguishing particular 2-way problems. For example, the HTR feature achieves 92% in classifying ND from FMD, whereas HMSF gives 83% in classifying MCI from FMD. Moreover, people with ND turn their heads more than those with FMD, and the accompanying person's presence plays a crucial role, as shown in Figure 6.2.

Table 6.2 Classification accuracy in percentage (%) of 3-way and 2-way classification tasks for the IVA_{18} dataset measuring the system performance using individual features with different classifiers (superscripts indicate the classifier: $L-SVM^1$, $RBF-SVM^2$, kNN^3, LR^4 and DT^5).

Feature	**No. of features**	**ND/MCI/FMD**	**ND/MCI**	**ND/FMD**	**MCI/FMD**
HTR	1	56^2	$\mathbf{67}^2$	$\mathbf{92}^2$	$58^{1,3,4}$
HTR+HTSF	13	$\mathbf{67}^2$	50^3	$75^{1,2,3}$	67^2
HMSF	54	50^2	$\mathbf{67}^2$	58^3	$\mathbf{83}^{1,4}$

6.5.2 Statistical Analysis

Table 6.3 shows the resulting features of the two tests, T-test and Wilcoxon, that may help to understand the most significant features that can distinguish between two groups. For the ND/MCI group, the t-test shows that the SD and Var of the yaw angle and velocity are statistically significant. This means that the participants with ND significantly vary the yaw angle and velocity due to long head turns and movement. For the ND/FMD group, only the SD of the yaw angle is considered statically significant to distinguish between the two groups. This indicates that participants with ND have longer head turns and movement than those in the MCI and FMD groups.

Table 6.3 Significant T-test and Wilcoxon test results of HMSF features (T=T-test, W=Wilcoxon test).

Classes	**Feature**	**P-value (T/W)**
ND/MCI	SD of yaw angle	.04 (T)
	Var of yaw velocity	.05 (T)
ND/FMD	SD of yaw angle	.04 (T)
MCI/FMD	Min. of pitch angle	.02 (W)
	Min. of pitch velocity	.02 (W)
	Mean of pitch velocity	.04 (T)
	Mean of pitch angle	.05 (T)

In the MCI/FMD group, the mean pitch angle and velocity are statistically significant. This shows that participants with MCI have a lower mean value than those with FMD after these features are observed. This suggests that MCI participants tend to look down more and show slower head movement than FMD participants. On the other hand, the 32 features tested using a Wilcoxon test resulted in only two features, minimum pitch angle and velocity, which are statistically significant and indicate that participants with MCI move their heads down a lot.

6.5.3 Effect of Feature Fusion

Feature fusion performance is measured by combining: 1) EBR+HTR, 2) EBR+HTR+HTSF, 3) EBR+HMSF, and 4) fusion of all features. The fusion of EBR with other features is carried out by investigating the performance using the SM and different thresholds: 7, 70, and 700. Table 6.4 provides the experimental results. It is clear that the feature fusion of EBR+HTR and EBR+HTR+HTSF improves the performance in differentiating ND and MCI with an accuracy of 75%, compared with 67% when all the visual features are combined. In addition, the combinations of EBR+HTR+HTSF and EBR+HMSF show the best results obtained for the MCI/FMD groups with 92% compared with the performance of individual features at 67% in HTR+HTSF and 83% in HMSF. The performance of combining EBR+HTR+HTSF gives 92% in distinguishing ND and FMD. ND/MCI can be classified with 75% using any fusion type except the fusing of EBR+HMSF features. It can be seen that feature fusion mainly helps in differentiating 2-way problems (ND/MCI and MCI/FMD), which is an indicator of the importance of these hand-crafted visual features, especially the EBR, HTR and HTSF, to capture the differences between these groups.

Table 6.4 Classification accuracy in percentage (%) of 3-way and 2-way classification tasks for the IVA_{18} dataset measuring the system performance using feature fusion, and feature selection with different classifiers (superscripts indicate the classifier: $L-SVM^1$, $RBF-SVM^2$, kNN^3, LR^4 and DT^5).

Feature	No. of features	ND/MCI/FMD	ND/MCI	ND/FMD	MCI/FMD
EBR+HTR	8	**78**^{5}	75^{1,4}	75^{2,5}	58^{3}
Feature selection	3	61^{1,4,5}	50^{1,2,4,5}	83^{1,4}	**92**^{1,2,4}
EBR+HTR+HTSF	83	72^{2}	75^{2}	**92**^{1,2}	**92**^{2}
Feature selection	40	**78**^{2}	**83**^{2}	83^{1,2,3,4}	**92**^{2}
EBR+HMSF	124	67^{2}	67^{2}	83^{2}	**92**^{2}
Feature selection	61	67^{2}	67^{2}	83^{1,2,4}	**92**^{1,2,4}
EBR+HTR+HTSF+HMSF	137	72^{1}	75^{1,2,4}	83^{4}	83^{1,2,4}
Feature selection	116	72^{1}	75^{1,2,4}	83^{4}	83^{1,4}

6.5.4 Effect of Feature Selection

Feature selection is then performed using the RFECV approach, which is applied to every feature fusion, as shown in Table 6.4. The table shows that feature selection for the fusion of EBR+HTR+HTSF improves the 3-way classification accuracy from 72% to 78%. Moreover, the feature selection of feature fusion for EBR+HTR gives better performance in some 2-way problems, such as ND/FMD and MCI/FMD, compared with their performance before the feature selection. For example, the classification of ND/FMD (and MCI/FMD) groups improved from 75% to 83% and from 58% to 92%, respectively, using only EBR with HTR. However, using HTR and HTSF with EBR enhances the accuracy from 75% to 83% for ND/FMD groups, respectively. Feature selection shows some improvements in the performance, but it was not significant compared to their performance without feature selection. This shows that each feature has its contribution to the classification. However, the effect of feature selection is significant compared with the performance of individual features, such as HTR, HTR+HTSF and HMSF.

6.5.5 Effect of Low-Level Features

A GMM is used to transform the LLFs' sequences into equal lengths. To do this, it is important to know how many clusters are needed, but firstly, it is important to use a test to assess whether the cluster number is the optimal one or not. The Bayesian information criterion (BIC) is therefore used to determine the optimal number of clusters. This criterion is a built-in function in the estimator of Scikit-Learn's GMM. It is used to select a model from a finite set of models and is based on the likelihood function. The likelihood function can be increased by adding parameters, which can lead to overfitting. The BIC addresses this by proposing a penalty term to test the models. If the BIC value is lower, the penalty is lower, hence a better model.

After the parameters have been tuned, the optimal number of components will correspond to the minimum value of the BIC score. The BIC, in this case, shows that the optimal number of components is 24 components (see Figure 6.9). Then, a GMM model is built for each participant using 24 components, resulting in (9×24) as the feature dimension. After that, the models are given to several supervised machine learning classifiers. The results obtained are presented in Table 6.5. The table shows that the GMM gives accuracy scores above the chance-level and the highest accuracy score is 74% for classifying ND/FMD groups from each other. The features produced from GMM does not show any significant increase in accuracy compared to previous features: HTR, HTSF and HMSF. The difference between the GMM obtained results and the highest accuracy score achieved in the feature selection of EBR+HTR+HTSF in Table 6.4 is statistically significant ($p = 0.02$).

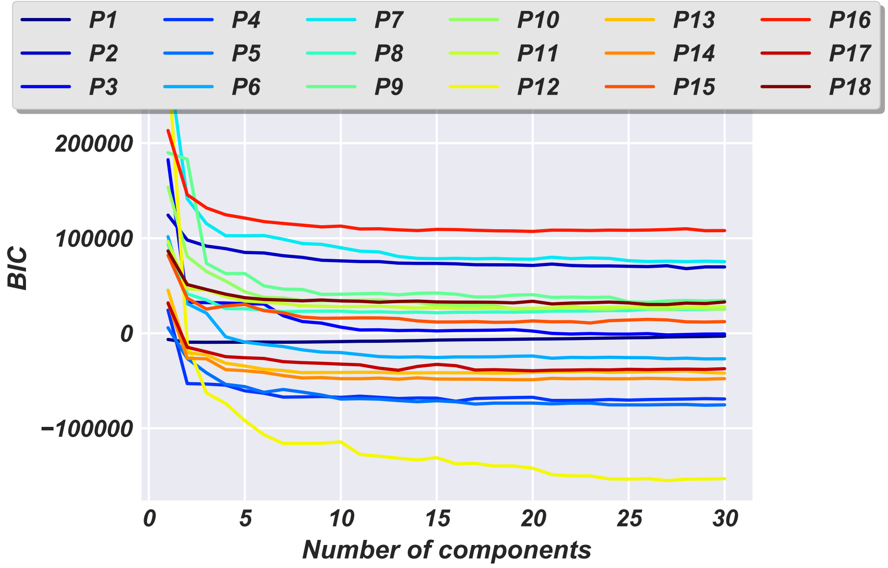


Fig. 6.9 The Bayesian information criterion (BIC) scores for GMM using different numbers of components.

Table 6.5 Classification accuracy in percentage (%) of the 3-way and 2-way classification tasks for the IVA_{18} dataset measuring the system performance using the means and the variance of 24 GMM components for the LLFs. (Superscripts indicate the classifier: $L-SVM^1$, $RBF-SVM^2$, kNN^3, LR^4 and DT^5).

Feature	No. of features	ND/MCI/FMD	ND/MCI	ND/FMD	MCI/FMD
LLFs	24	48^3	66^5	74^5	67^2

6.6 Evaluating the System on the IVA_{52} Dataset

As stated earlier, this chapter presents an evaluation of the head movement features on a larger dataset to distinguish between ND, MCI, FMD and HC. This section presents the classification results of the four-way, three-way and the combination of two-way tasks. Performance is measured using individual features HTR, HTSF and HMSF.

In this experiment, a three-way classification task (ND vs. MCI vs. FMD) and several two-way classification tasks (ND vs. MCI, ND vs. FMD and MCI vs. FMD) are carried out using the same four classifiers used when testing IVA_{18} in order to enable a fair comparison between the results obtained with the results reported in previous section. Then, a four-way classification task and more combinations of two-way classification tasks are explored. Four metrics are calculated for each classifier: accuracy, precision, recall and F-measure.

6.6.1 Three-way and Two-way Classification Tasks

Table 6.6 presents the classification accuracy results for three-way and two-way classification tasks. It can be seen that the highest results obtained are for the HTR+HTSF features, at 59% for the three-way classification problem and 92% for ND vs. FMD. Moreover, the use of HMSF features achieves 62% and 90% accuracy in distinguishing ND from MCI and MCI from FMD, respectively. A comparison is then carried out between the results obtained using the IVA_{52} dataset and the IVA_{18} results, reported in Section 6.5. Table 6.7 shows the results achieved in both datasets using the same classifiers as used in the IVA_{18} dataset. In general, it is clear from the table that the performance decreases when the IVA_{52} dataset is used. However, the HTR+HTSF feature gave the highest accuracy results for the three-way classification problems in both datasets. Some features are better at classifying particular two-way problems. For instance, the HTR+HTSF feature achieves 92% accuracy in classifying ND from FMD using the IVA_{52} dataset. In addition, the HTR feature gave 72% accuracy in differentiating MCI from FMD. These results confirm the findings set out in Chapter 6.5 that people with ND tend to turn their heads more than those with FMD, resulting from the presence of an accompanying person.

6.6.2 Four-way and Two-way Classification Tasks

As stated earlier, the head movement features are evaluated in two classification tasks: four-way classification problems (ND vs. MCI vs. FMD vs. HC) and two-way classification problems, including ND vs. HC, MCI vs. HC and FMD vs. HC. Further, two-way classification problems are then investigated to differentiate participants with memory problems (MP) from HC. Then, the data is divided into an HC group and an MP group, consisting of ND, MCI and FMD. Another classification task is then conducted to classify participants with dementia from HC and FMD. The data is, therefore, divided into the dementia group (D) and the non-dementia group (Non-D). The D includes ND and MCI and the Non-D includes FMD and HC. Table 6.8 presents the results of the four-way and two-way classification problems. It is clear from the table that HTR and HTR+HTSF, which is the derivative features from the HTR, are considered informative features. It is also shown that classifying FMD from HC is the most challenging classification problem.

Table 6.6 Classification accuracy of three-way and two-way classification tasks for the IVA_{52} dataset measuring the system performance using individual features with the KNN classifier. The number of features is indicated between the parentheses ().

Classification Task	Feature	Accuracy	Precision	Recall	F-measure
ND/MCI/FMD	HTR (1)	45%	31%	46%	37%
	HTR+HTSF (13)	**59%**	**60%**	**58%**	**58%**
	HMSF (54)	53%	59%	55%	53%
ND/MCI	HTR (1)	53%	51%	51%	46%
	HTR+HTSF (13)	52%	52%	50%	**69%**
	HMSF (54)	**62%**	**62%**	**62%**	62%
ND/FMD	HTR (1)	73%	76%	70%	71%
	HTR+HTSF (13)	**92%**	**89%**	**89%**	**89%**
	HMSF (54)	69%	68%	66%	66%
MCI/FMD	HTR (1)	71%	72%	73%	72%
	HTR+HTSF (13)	67%	71%	69%	66%
	HMSF (54)	**90%**	**92%**	**88%**	**88%**

Table 6.7 Classification accuracy in percentages (%) of the three-way and two-way classification tasks for the IVA_{18} and the IVA_{52} datasets measuring the system performance using individual features with different classifiers (superscripts indicate the classifier: $L-SVM^1$, $RBF-SVM^2$, kNN^3, LR^4 and DT^5).

Dataset	Feature	ND/MCI/FMD	ND/MCI	ND/FMD	MCI/FMD
IVA_{18}	HTR	56^2	$\mathbf{67}^2$	$\mathbf{92}^2$	58^3
	HTR+HTSF	$\mathbf{67}^2$	50^3	75^3	67^2
	HMSF	50^2	$\mathbf{67}^2$	58^3	$\mathbf{83}^1$
IVA_{52}	HTR	38^2	$\mathbf{52}^2$	56^2	$\mathbf{72}^3$
	HTR+HTSF	$\mathbf{59}^2$	$\mathbf{52}^3$	$\mathbf{92}^3$	63^2
	HMSF	34^2	48^2	69^3	53^1

6.7 Discussion

The findings show the importance of visual cues, such as the head movement and LLFs features, in differentiating between ND, MCI and FMD with 78% accuracy. The feature

Table 6.8 Classification accuracy of four-way and two-way classification tasks for the IVA_{52} dataset measuring the system performance using individual features with the KNN classifier. The number of features is indicated between the parentheses.

Classification Task	**Feature**	**Accuracy**	**Precision**	**Recall**	**F-measure**
ND/MCI/FMD/HC	HTR (1)	45%	46%	42%	**59%**
	HTR+HTSF (13)	**48%**	**50%**	**46%**	53%
	HMSF (54)	44%	45%	44%	44%
MP/HC	HTR (1)	72%	71%	71%	71%
	HTR+HTSF (13)	64%	63%	62%	62%
	HMSF (54)	**73%**	**73%**	**72%**	**72%**
D/Non-D	HTR (1)	69%	66%	64%	64%
	HTR+HTSF (13)	**75%**	**72%**	**71%**	**72%**
	HMSF (54)	73%	70%	70%	70%
ND/HC	HTR (1)	**83%**	**85%**	**84%**	**85%**
	HTR+HTSF (13)	72%	77%	73%	73%
	HMSF (54)	74%	75%	74%	74%
MCI/HC	HTR (1)	**74%**	**74%**	**74%**	**74%**
	HTR+HTSF (13)	69%	68%	68%	68%
	HMSF (54)	69%	68%	68%	68%
FMD/HC	HTR (1)	**67%**	**65%**	**64%**	**64%**
	HTR+HTSF (13)	45%	25%	44%	32%
	HMSF (54)	58%	59%	58%	58%

selection of the HTR+HTSF combined with EBR is the only case in which the performance shows an improvement. However, other feature types of feature fusion show improvements in the performance of particular classes. A possible explanation for this could be that every feature has its contribution in differentiating the three classes (Bishop and Nasrabadi, 2006).

Figure 6.10 shows that confusion mostly happens in the MCI and ND groups. Two participants from the MCI group are incorrectly classified as FMD because neither of them is moving their heads a lot. Interestingly, the same two participants are predicted as ND from their EBR. Although the ND participants are correctly classified using the EBR feature, two ND participants are identified incorrectly as MCI using the feature fusion. This rather

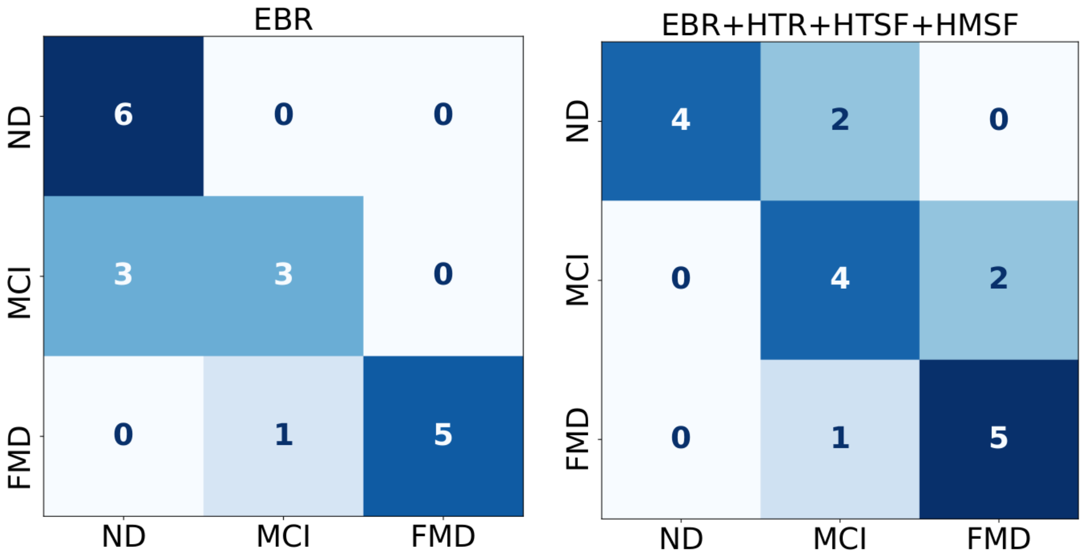


Fig. 6.10 Comparison in terms of the performance between the eye blink rate (EBR) and the feature fusion of all features (EBR+HTR+HTSF+HMSF) for the IVA_{18} dataset using a confusion matrix (rows: true labels and columns: predicted labels).

contradictory result may be because one does not have an accompanying partner, and the other has a partner but did not turn his head often.

These possible sources of error could have been caused by the lack of diagnostic details about what dementia type the participants have. This information could help in error analysis because people with some particular types of dementia may not show a head turn cue, such as VaD and DLB (Fukui et al., 2011). This information is essential for developing an automatic tool to handle the overlapping diseases, such as AD, VaD, DLB, and behavioural variant frontotemporal dementia (FTD).

Looking more closely at the interaction between gender, age and whether a partner is present is of interest. Figure 6.11 shows the number of participants who attended with a partner and those who attended alone according to group, gender and age. Larner (2012, 2018) and Tyson et al. (2019) suggested that ‘attending with’ is an additional cue to the head turn due to its effect on the latter. In accordance with the previous work results, Figures B.5, B.6, and B.7 show that participants who came with a partner generally show significant head turns and movements, which is consistent with related work findings (Durães et al., 2018; Larner, 2012). Moreover, they suggested that the presence of head turn indicates CI and AD. In contrast, Fukui et al. (2011) found that the presence of the head turn cue indicates CI whether he/she attended with a partner or came alone. The findings of this research, while preliminary, seem to be consistent with other research, which found that increases in EBR and head turn or movements can indicate a higher risk of progression to AD (Durães et al., 2018; Ladas et al., 2014; Larner, 2018). It is also assumed that gender and age may play a vital role in the HTR.

In the IVA_{18} data, all of the female participants came with a partner regardless of their diagnostic class, whilst only two men from the ND class came with partners, as shown in Figure 6.11. Moreover, in the IVA_{34} data, all the female participants who came with partners have health conditions, but only two males from MCI and HC classes came with a partner. Fukui et al. (2011) investigated the severity and the incidence of the head turn cue on 125 patients by observing whether the patients showed a head turn cue during a cognitive test, which is the revised Hasegawa Dementia Rating Scale, which takes about 10 minutes, and found that the females tended to bring a partner. Their findings showed that women find it easier to depend on someone else when they face difficulties, whereas men feel obligated to deal with difficulties without help. Previous work compared men and women with CI in terms of the prevalence of behavioural symptoms and found that 'help-seeking' and depression are more frequent in women (Lövheim et al., 2009). However, men showed more regressive and aggressive behaviours than women. Those findings contradict those of previous work that found the partner presence is independent of the gender (Holland and Larner, 2013; Larner, 2005) but dependent on the age (Larner, 2014a). Moreover, Larner (2005) reported that the head turn indicates a CI regardless of gender or age. However, there is no information on whether these factors are affected by other external factors such as culture.

These findings are in accord with those of recent studies indicating that head movement can provide rich information in differentiating ND, MCI, FMD and HC. This may be helpful in gauging the severity of the CI and whether participants have a higher risk of progression to AD. However, caution must be applied with a small sample size, such as the dataset used in this research, as the findings might not be generalisable.

6.8 Summary

This chapter has investigated the feasibility of automatically detecting ND, MCI and FMD using the head movement features on the IVA_{18} and then evaluated the system on a larger dataset (IVA_{52}) with many variations in the recording settings and devices used. This investigation involved analysing the participants' behaviour and the challenges observed in the head movement data. Moreover, the performance for the IVA_{18} dataset was measured using each feature individually, the fusion of these head features with the EBR feature, and then when the features were selected, and the LLFs. Then, the system was evaluated on the IVA_{52} dataset to differentiate ND, MCI, FMD and HC groups by four-way, three-way and a combination of two-way problems. Finally, the findings have been analysed and compared with related work.

For the IVA_{18} dataset, the findings showed that identifying ND/MCI and MCI/FMD using the HTR was challenging, with an accuracy of 67%. These classification problems are challenging cases even in the clinic. However, the classification of ND/FMD using the same feature achieved high performance, with an accuracy of 92%. The HMSF identified the MCI/FMD better than the HTR feature with an accuracy of 83%. The reason for this is that the features of the three angles were very informative. The results also showed how the fusion of these visual features is significant for classifying ND, MCI and FMD with an accuracy of 78%. It also showed that the LLFs (frame-by-frame) features did not enhance the performance for the 3-way and 2-way classification problems compared with other features.

It was found that the same features, which gave the highest results on the IVA_{18} dataset, obtained the highest accuracy results for the three-way problem and the two-way ND vs. MCI problem in the IVA_{52} datasets. In addition, the features of HTR and HTR+HTSF helped to perform better in the two-way classification problems ND vs. FMD and MCI vs. FMD. The two-way classification problem MCI vs. FMD showed a decrease in performance after increasing the size of the dataset using the HMSF feature. However, it showed increased accuracy when the HTR feature was employed. The four-way and two-way classification problems also showed good accuracy results, and FMD vs. HC was a challenging task.

The generalisability of these results is subject to certain limitations, for instance, the small dataset size and the lack of some diagnostic information, such as the type of MCI or ND that the participants have. Despite these limitations, investigating these visual features in the detection of CI lays the groundwork for future research into CI detection based on specific visual features using in-the-wild data to build a reliable system instead of having high performance in lab-controlled data that does not reflect real-life data. The next chapter will validate the work in this chapter and Chapter 4 on a public dataset to compare the results achieved with related work.

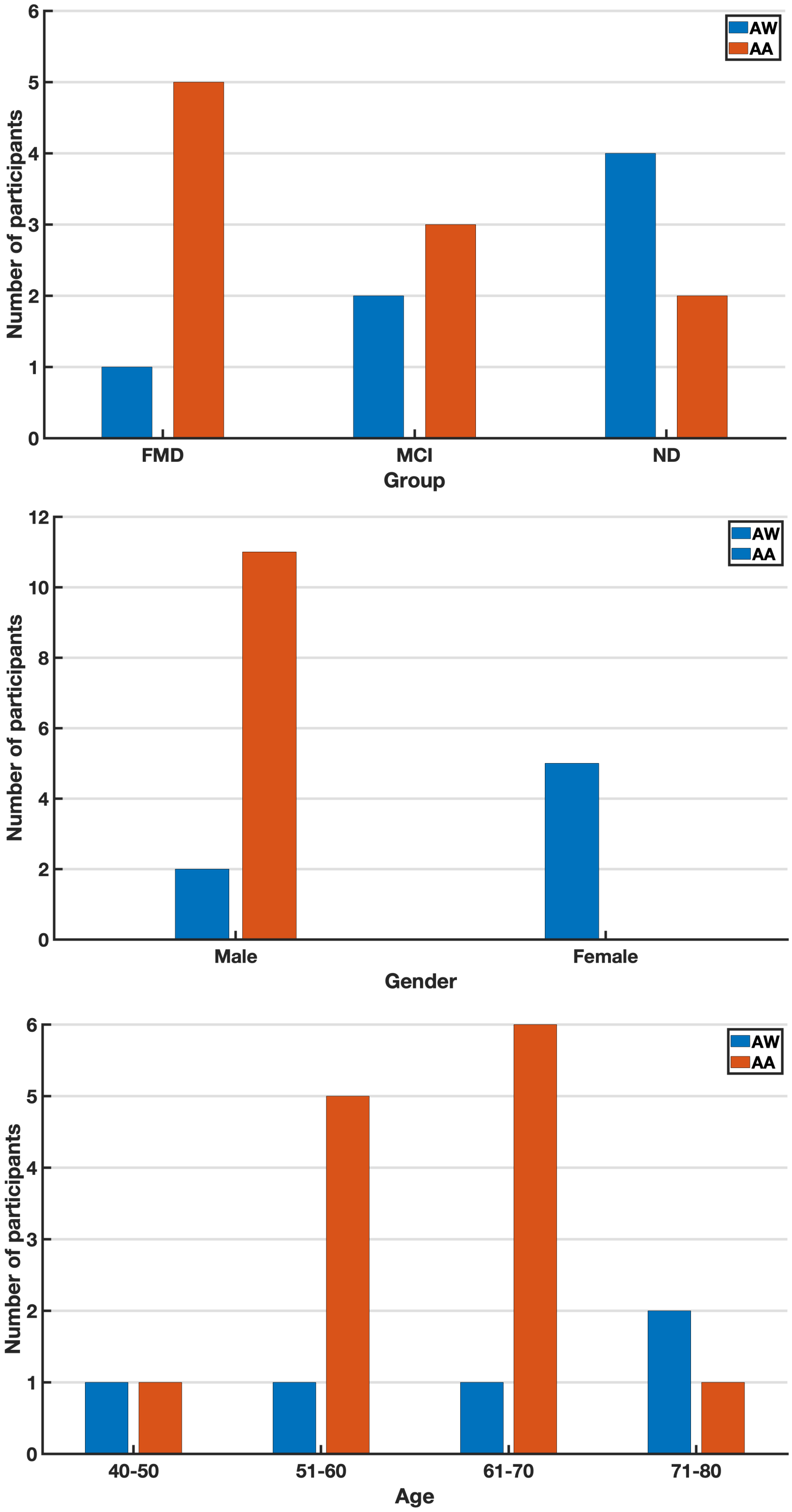


Fig. 6.11 Analysis of the HTR based on participants' group, gender, and age (AW: Attend with, and AA: attend alone).

Chapter 7

Towards the Automatic Classification and Regression Analysis of Depression

"And sometimes I have kept my feelings to myself
because I could find no language to describe them in."
— Marianne Dashwood

7.1 Introduction

In the previous chapters, the system for detecting cognitive impairment was introduced and evaluated on data recorded in the wild for people with cognitive impairment (CI) and related conditions. Despite the issues caused by the nature of the data recording settings (i.e., in-the-wild scenarios), a robust feature extraction method was developed to overcome these challenges. This method was used to extract visual features, such as eye blink rate (EBR), head turn rate (HTR), head turn statistical features (HTSF) and head movement statistical features (HMSF) and then feed them to the classifiers. The performance obtained showed promising results (see Chapters 4 and 6).

This chapter will explore the performance of the system by applying it to a widely used depression dataset and will then compare its performance with those reported in related studies, not to compete with state-of-the-art approaches (Jan et al., 2014; Kaya et al., 2014; Pérez Espinosa et al., 2014; Valstar et al., 2014). The dataset comprises two tasks: classification, which distinguishes between depressed and non-depressed people, and regression, which predicts depression severity. These tasks are appropriate for the aim of this current study, which is generalisability within individual datasets because the IVA_{18} dataset and the depression dataset differ in several factors that may impact the generalisation. These factors are the difference in the data collection procedure, the duration of the session for each participant in each dataset, the hardware used and the recording environments. These experiments could give insight into the efficiency of the approaches described in previous Chapters 4 and 6, which involves evaluating the performance using the same feature sets and classifiers.

Since depression might affect people's behaviour, a considerable amount of literature has been published on the automatic detection of depression using visual behaviour (He et al., 2022; Pampouchidou et al., 2017; Ringeval et al., 2019). A number of techniques have been developed to detect depression using facial cues, such as EBR (Alghowinem et al., 2015; Gupta et al., 2014; Zhou et al., 2015) and head movement analysis (Alghowinem et al., 2020; Morency et al., 2015; Ringeval et al., 2018, 2019; Zhou et al., 2015) due to the strong correlation between these facial cues and depression. People with depression may not exhibit help-seeking behaviour for several reasons, such as fear of the associated stigma, embarrassment, and a preference for self-reliance (Gulliver et al., 2010). Other factors that could cause that behaviour are the cost, location and service availability (Draucker, 2005; Kuwabara et al., 2007; McCann and Lubman, 2012). Moreover, problems related to not seeking help include a lack of awareness of the benefits of seeking help and an unwillingness to express emotion (Gulliver et al., 2010) because people with depression can tend to avoid communicating with others and telling them what they are going through because they believe that nobody can understand them or because they do not know how to express what they feel (Ellgring, 2007). Employing visual-based approaches can benefit a doctor's diagnosis by providing an objective diagnostic aid system because estimating the severity of someone's depression can often be difficult (Nemeroff, 2007). In addition, a visual-based approach might be considered language-independent (Gogate et al., 2020; Papakostas et al., 2017; Pusdekar and Chhaware, 2014), which is helpful if the participants show an unwillingness to talk about themselves due to one of the factors listed above.

The rest of this chapter is organised into three main sections. Section 7.2 describes and analyses the dataset. Section 7.3 presents the results and then discusses them. Section 7.4 then presents the summary of the chapter.

7.2 Experiments

This section outlines the data and presents all the details regarding the features used in the experiments. The extracted features are then analysed. The classifiers and regressors used are then described and, finally, the feature selection is described.

7.2.1 Data

The videos used are from the audio-visual emotion challenge 2014 (AVEC 2014) dataset (Valstar et al., 2014). It consists of 300 videos recorded individually of German speakers using a webcam and microphone. The participants completed two tasks: answering questions (Freeform task) and reading a paragraph aloud (Northwind task). For each task, a total of 150 videos were recorded, ranging from six seconds to four minutes and eight seconds in duration. The sessions took place in several quiet settings. Some participants made more than one recording, with two weeks between the recordings. There is only one participant in each video clip. The age of the participants ranges from 18 to 63 years old; the average is 31.5 years old.

Every recording was labelled with a single score for the severity level of the depression based on the BDI-II scale (Beck et al., 1996) and a self-report consisting of 12 question-and-answer scores from 0 to 3, with the total score ranging from 0 to 63. This gives four categories of depression, as presented in Table 7.1. Figure 7.1 shows an example of one participant with different levels of depression severity, which are recorded in challenging scenarios using a webcam. The reason for having one participant with such different levels of depression could be the different recording times. In this chapter, only the Freeform data with 150 videos is used as Ebert et al. (1996) found no difference in the EBR between depressed and non-depressed people in the reading task. In addition, Alghowinem et al. (2013a) showed that depressed and non-depressed people had almost the same EBR. However, a difference was found in the eye blink duration when answering a question task. That is why only the Freeform task data is used in this study. The distribution of the levels of depression severity in the AVEC 2014 data (Freeform task) is presented in Figure 7.2. The organizers of the AVEC 2014 challenge divided the data equally into train, development and test partitions.

Table 7.1 The score range of each depression level.

Score	**Range**
0 - 13	Minimal
14 - 19	Mild
20 - 28	Moderate
29 - 63	Severe

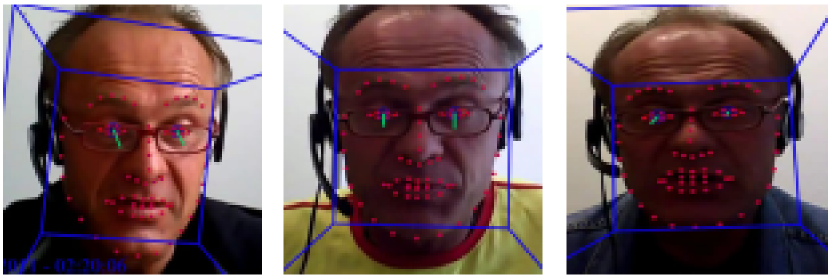


Fig. 7.1 Examples of facial images for one participant with different depression levels (BDI–II scores of 6, 19, and 33 from left to right) from AVEC 2014 Valstar et al. (2014)

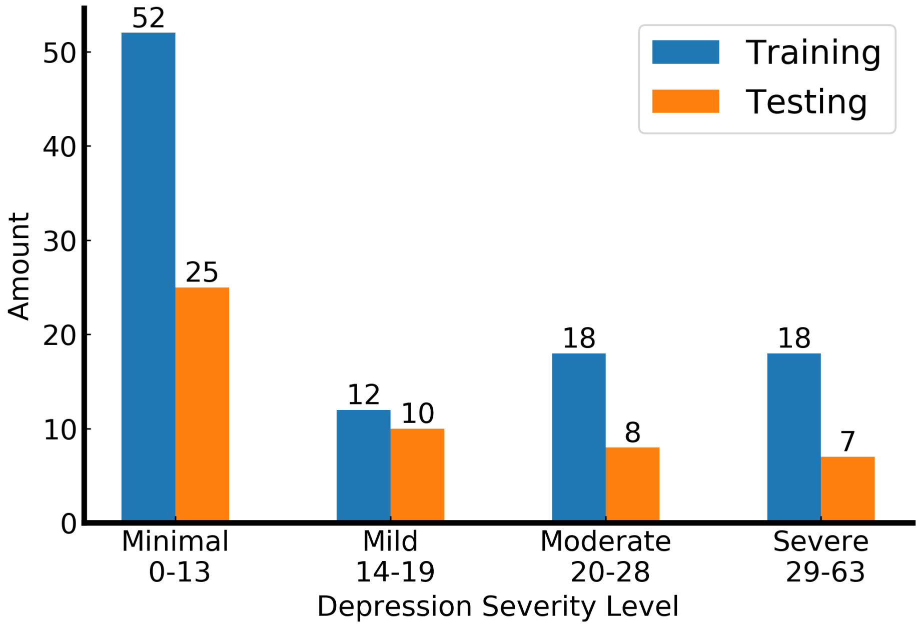


Fig. 7.2 The distribution of the depression severity level for the training partition of the Freeform task in the AVEC 2014 dataset.

7.2.2 Analysing the Signal

Prior to the detection of facial landmarks, the AVEC 2014 data does not require any preprocessing operation, unlike the IVA_{18} data. Therefore, the first step to perform is the facial landmarks prediction for each video frame. This is implemented automatically using the OpenFace toolkit[1]. From the output of the OpenFace, the eye aspect ratio (EAR) is computed for both eyes' landmarks for each frame. The average of both eyes' EAR is used (see Chapter 4 for more detail). In addition, the estimated three angles, pitch, yaw and roll, of the head pose are used to calculate the HTR, HTSF and HMSF features, as described

[1] https://github.com/TadasBaltrusaitis/OpenFace/

in Chapter 6. Three of the 150 videos needed some pre-processing operations due to the participant and face detector issues. For example, Figure 7.3 shows when the participant left the chair and led to a line of one constant value, which is a repeated value from the last time the face was detected. Then, a head turn to the left led to a high value of the EAR of 4.22, which confirms the observations in Chapter 4 when it is found that extremely high values are assumed to be head turns. This issue is fixed by removing this line of repeated values when the participant was not there. All features are normalised to have zero mean and unit variance. The following sub-sections present an analysis of the calculated EAR and the head movements.

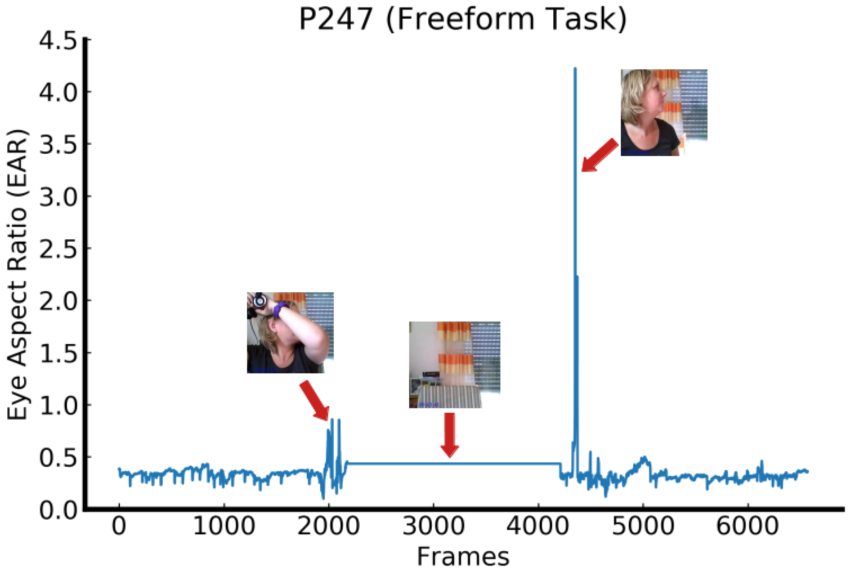


Fig. 7.3 The calculated EAR for participant P247 in the Freeform task, which shows the states where the participant left the chair and turned her head.

Analysing the Calculated EAR

Figure 7.4 shows the calculated EAR for two participants in the AVEC 2014 dataset. It also shows that the EAR is challenging, and the number of blinks cannot be inspected without seeing the videos. This leads to observing the data visually for all of the clips and determining the percentage of the challenging data for the eye blink detection task based on the number of challenging videos and the total number of videos. After observing the data for the Freeform task, determining whether the drifts represent blinks or not can be difficult without watching the video to check. Around 81.33% of the data is considered challenging. Thus, the multiple thresholds (MTs) approach could be used for such data, as described in Chapter 4.

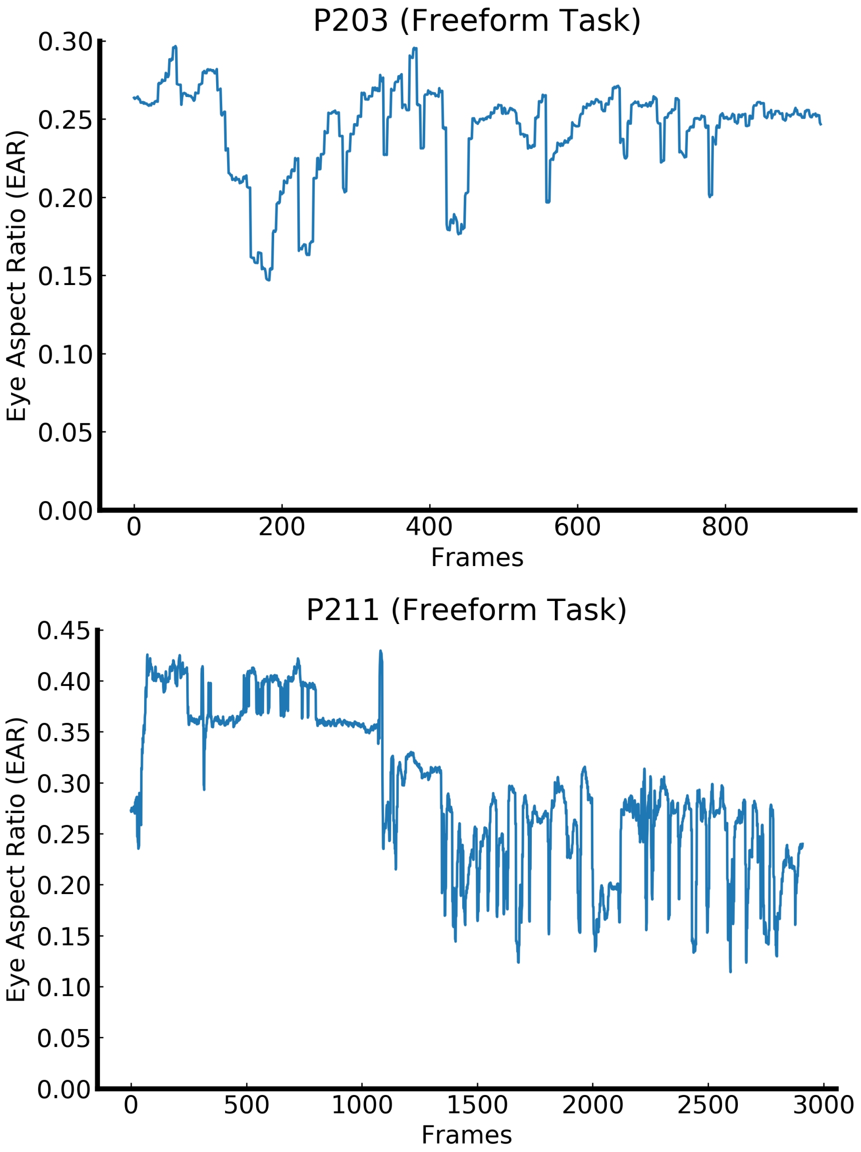


Fig. 7.4 Examples of the calculated EAR for two participants in the AVEC 2014 data doing the Freeform task.

Analysing the Estimated Angles of the Head Movement

This section analyses the Euler angles of the head movements for the AVEC 2014 data to show the participants' different behaviour when doing the Freeform task (i.e., answering questions). It also shows an additional motive to the previously mentioned ones to exclude the Northwind data. Figure 7.5 shows line plots of the pitch, yaw and roll angles, indicating the head movements throughout the videos for three participants. These three participants were chosen to illustrate the variety of participants' behaviour during this task. The figure shows that the participants exhibit head movements and turns when performing the Freeform task. However, the average of the angles of the head movements in the Freeform data is significantly lower than in the IVA_{18} data.

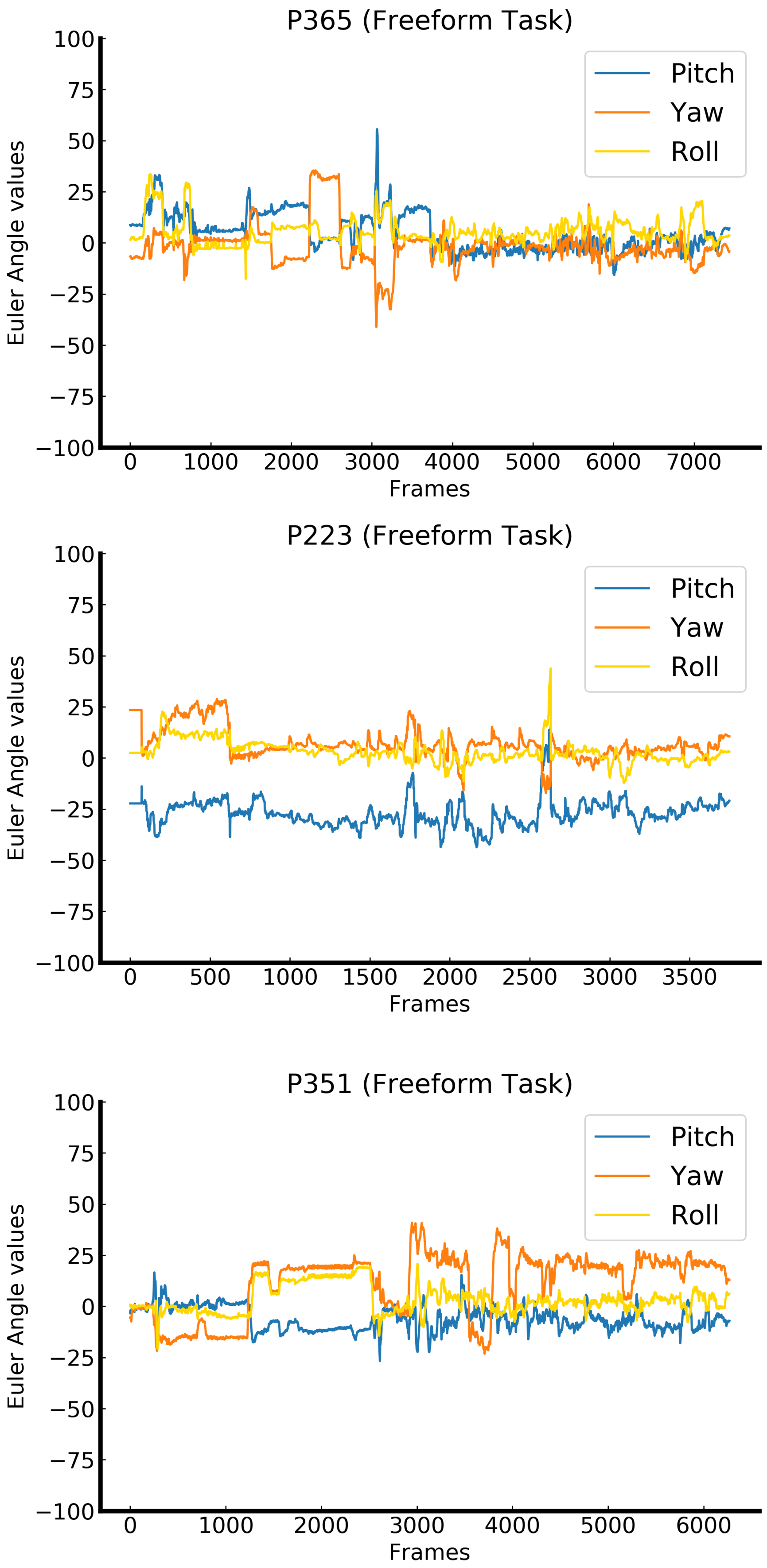
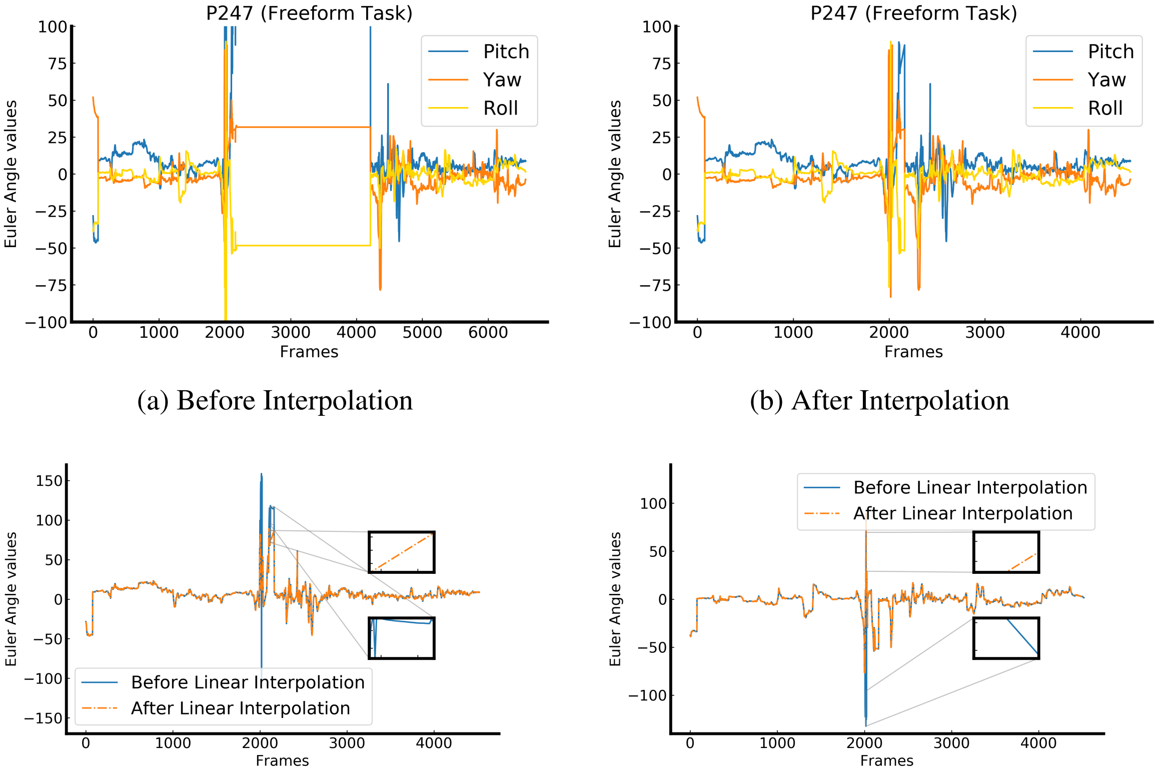


Fig. 7.5 Examples of the calculated three angles for the Freeform task.

The same issues were observed for the three participants from the Freeform data, as described above, which exist in the head movements, which can be seen in Figure 7.6. Observing this data shows that the three participants from the Freeform data present a problem related to the estimated head pose angles. This problem is addressed by automatically removing repeated or zero values, as described above. Another problem related to the estimated angles is the extremely high values that resulted from either the head movements angle being difficult to detect even by OpenFace or when the participant left the chair. This problem is resolved by using linear interpolation where the outlier values appear, as illustrated in Figure 7.6.



(a) Before Interpolation

(b) After Interpolation

(c) Pitch Angle

(d) Roll Angle

Fig. 7.6 An example of the original signal of the head movements and after the linear interpolation operation for roll and pitch angles for a participant.

7.2.3 Classification

As stated previously, two tasks are explored: classification for detecting depressed from non-depressed people and regression for determining the severity of depression. This section describes the classification task, and the next section presents the regression task. For the classification, several machine learning classifiers are used to measure the performance: support vector machine with Linear (L-SVM) and RBF kernels (RBF-SVM), logistic regression (LR), k-nearest neighbours (KNN) and decision trees (DT). The hyper-parameter values are set based on a grid search, and the rest of the parameters are set to the default values (Pedregosa et al., 2011). The data in this task is divided into two groups: the non-depressed group (NDG) and the depressed group (DG), where the threshold used for the BDI-II here is 13, which means that any participant with a score above 13 is depressed, as predefined in Table 7.1. Table 7.2 shows the number of participants in each group based on the task and the partition. The test set is held out consisting of 50 video clips (NDG:25 and DG:25), whereas the two remaining sets, train and development sets, are used individually and combined as a training set, including 50 and 100 video clips, respectively. The two sets are combined to increase the training set. For the evaluation, accuracy, precision, recall, and F-measure metrics are calculated for each classifier. The accuracy metric is adopted because the test set contains balanced data.

Table 7.2 Presenting the number of participants in each group for the classification task (NDG: non-depressed group, DG: depressed group).

Task	**Partition**	**Number of NDG**	**Number of DG**
FreeForm	Training	26	24
	Development	26	24
	Testing	25	25

7.2.4 Regression

The regression task is conducted to predict the score of depression severity. To do this, three regressors are used, support vector regression with both kernels, Linear (L-SVR) and RBF (RBF-SVR), and KNN with both weights and DT. These regressors are for evaluating the performance and comparing it with related work. The train and test sets are split, as described previously in the classification task. Following related work, two metrics are used to measure the prediction of depression severity, which are the mean absolute error (MAE) and the root mean squared error (RMSE). Details about these equations are described in Section 2.5

7.2.5 Feature Selection

The recursive feature elimination cross-validation (RFECV) algorithm is adopted to improve the performance of the classification and regression results. This algorithm is popular due to its simplicity of use and configuration. It is known that RFECV is very efficient in selecting the optimal features in the training data, as described in Section 6.3.6. The number of features to select is a significant hyper-parameter to the performance. This could, therefore, be done automatically by RFECV via cross-validation evaluation for different feature numbers, which will then select the optimal number of features with the best mean score. The base algorithm can be evaluated to find which model gives the best result. For this study, the model with the best mean result is chosen for the feature selection. This process is applied to the classification and regression tasks.

7.3 Results and Discussion

As stated previously, this chapter aims to validate the method used in this research and enable comparison with related work. This section presents the results of the experiments for the two tasks: classification and regression. The performance achieved for each task is compared with related work.

7.3.1 Depression Detection

The performance of depression classification is measured for each feature individually, with feature fusion and when applying feature selection.

Feature Extraction and Classifiers

As stated earlier, the MTs approach calculates the thresholds based on the participants' overall minimums and maximums. Due to the high extreme values, the third standard deviation (SD) is used to remove them. After using the new range from the SD, different feature numbers based on the thresholds number of 5, 50, and 500 result in 5, 50 and 500 features using different state machine (SM) types. In addition, the HTR, the HTSF and the HMSF are also extracted. Then, they are fed to supervised machine learning classifiers. The classifiers used are the same used ones in the previous chapters and in related work (Alghowinem et al., 2020; Gupta et al., 2014; Zhou et al., 2015) to allow direct and fair comparison: SVM with linear kernel, LR, KNN with uniform weight and DT. The Python Scikit-learn package is used to train the models. The train and test sets are described in Section 7.2.3.

Results

Table 7.3 illustrates the best classification results for each feature based on different training sets: train, development and the combination of the train and development sets. The results show that the EBR feature using the MTs achieves the highest accuracy score in detecting depression, with 68% of accuracy when the train set and the combined train and development sets as a training set are used. They also show that the EBR gives the highest precision score of 71% when the train+development as a training set is used and gives the highest F-measure score of 68% when only the train set is used as the training set. In addition, the HMSF also achieved good performance at 64% when the training set was larger with 100 video clips. In contrast, the HTR feature, either alone or combined with HTSF, gives 50%, a chance level. This indicates that depressed people may not turn their heads to the left and right often, which is in line with (Alghowinem et al., 2013b).

Table 7.3 Classification results in percentage (%) of depression for the AVEC 2014 data on the test partition for testing each feature individually with different classifiers. (superscripts indicate the classifier: $L-SVM^1$, $RBF-SVM^2$, kNN^3, LR^4, DT^5, i: is the number of consecutive frames, Dev: Development).

Feature	**Training set**	**No. of features**	**Accuracy**	**Precision**	**Recall**	**F-measure**
EBR	Train	50 i=2:6	**68**^{1}	69	**68**	**68**
	Dev	5 i=2:4	56^{5}	56	56	56
	Train+Dev	5 i=2:6	**68**^{5}	**71**	68	67
HTR	Train	1	$50^{3,4,5}$	50	50	37
	Dev	1	$50^{1,2,4}$	50	50	33
	Train+Dev	1	$50^{2,3,4}$	50	50	42
HTR+HTSF	Train	13	$50^{3,4,5}$	50	50	37
	Dev	13	52^{1}	76	52	38
	Train+Dev	13	50^{3}	50	50	42
HMSF	Train	54	62^{3}	69	62	58
	Dev	54	62^{3}	61	60	59
	Train+Dev	54	64^{3}	64	64	64

The effect of carrying out feature fusion is then explored, and the results are shown in Table 7.4. The performance of feature fusion is measured by combining: 1) EBR + HTR, 2) EBR + HTR + HTSF, 3) EBR + HMSF and 4) all features. The results show that the fusion of the EBR with HTR and the fusion of the EBR and HTR+HTSF gives the highest accuracy, 68%. The fusion of all the features did not improve the performance of the previously mentioned results. Although the feature selection RFECV method is used to improve the performance, choosing the most contributed feature using feature selection did not help to improve the performance in detecting depression. The reason for this could be that the dimension of the combined feature vector (117 features) is not that large, and all the features contribute significantly to the classification task. Interestingly, the highest results are achieved when the eye blink duration is long, up to 26 consecutive frames below the threshold (i=2:26), shown in Table 7.4. A previous study by Alghowinem et al. (2013a) found that people with depression tend to have long blinks.

Table 7.4 Classification results in percentage (%) of depression for the AVEC 2014 data on the test partition for testing feature fusion and feature selection with different classifiers. (Superscripts indicate the classifier: $L-SVM^1$, $RBF-SVM^2$, kNN^3, LR^4, DT^5, i: is the number of consecutive frames, Dev: Development).

Feature	Training set	No. of features	Accuracy	Precision	Recall	F-measure
EBR+HTR	Train	51 i=2:26	**68**^{3}	68	**68**	**68**
	Dev	501 i=2:6	60^4	60	60	60
	Train+Dev	501 i=2:18	68^3	**69**	68	68
EBR+HTR+HTSF	Train	63 i=2:26	**68**^{3}	68	68	68
	Dev	513 i=2:6	58^3	58	58	58
	Train+Dev	513 i=2:16	58^4	61	58	55
EBR+HMSF	Train	554 i=2:12	64^4	67	64	63
	Dev	104 i=2:3	64^2	64	64	64
	Train+Dev	59 i=2:6	66^3	66	66	66
EBR+HTR+HTSF+HMSF	Train	567 i=2:12	66^4	68	66	65
	Dev	117 i=2:3	64^2	64	64	64
	Train+Dev	117 i=2:6	66^3	66	66	66
Selected features (RFECV)	Train+Dev	47 i=2:6	64^3	64	64	64

Comparing Performance with Related Work

A comparison between the performance achieved in this research and that in previous work for depression detection on the AVEC 2014 dataset is illustrated in Table 7.5. Accuracy and recall metrics are used because Senoussaoui et al. (2014) used the accuracy metric while Alghowinem et al. (2015) used the recall metric to measure the performance. The performance of the accuracy and recall metrics in the work described in this chapter is similar, so their performances could be used to compare with those reported in previous work. However, any comparison is complicated because previous work differs from the experiments described in this chapter in a number of important ways. For instance, Senoussaoui et al. (2014) used the development set to measure the performance of their approach, whereas this work used the test set. Senoussaoui et al. (2014) showed good results in differentiating depressed from non-depressed people with 82% accuracy using the development partition as a test set. This differs from the findings presented here that achieved about 68% accuracy using the test partition as a test set.

Table 7.5 Depression classification results compared to other methods on AVEC2014 and the number of features (in parentheses) (Dev: Development, Rec: recall metric).

Paper	**Classification Approach (No. of video clips)**	**Classifier**	**Feature (No. of features)**	**Accuracy (or as otherwise noted)**
Senoussaoui et al. (2014)	Dev (50)	RBF-SVM	Features, based on LGBP-TOP	82.0%
Alghowinem et al. (2015)	LOO Cross-validation (32)	RBF-SVM	Eye Activity (31)	Rec (81.3%)
			Head Pose (29)	Rec (68,8%)
			Feature fusion (60)	Rec (75.0%)
			Feature selection (12)	Rec (68.8%)
The proposed approach	Train-test (150)	L-SVM	EBR (50)	68.0%
		KNN	HTR+HTSF (13)	50.0%
			HMSF (54)	64.0%
			Feature fusion (117)	66.0%
			Feature selection (47)	64.0%

On the other hand, Alghowinem et al. (2015) used only 32 video clips from the Freeform task. This is because they were interested in analysing childhood storytelling to match their interviews from their dataset. In addition, they adopted the leave-one-out cross-validation approach for the classification task, which is different from the approach used in this chapter. Differences between the results reported by Alghowinem et al. (2015) and those in this thesis may have been influenced by the differences previously mentioned. However, both Alghowinem et al. (2015) and this thesis work share a number of key features. Firstly, both agree that eye features are significant and achieve the highest performance compared to head-related features. Moreover, the feature selection did not improve the performance of the classifier. Even though the hand-crafted features, which are used, are simple, they achieved encouraging results.

7.3.2 The Severity of Depression Detection

Besides detecting depression, estimating the severity of the depression is implemented by estimating the BDI-II score for each participant. In this task, the performance is measured for each feature individually, combined and when applying feature selection.

Feature Extraction and Regressors

The same features used in the classification task described previously are explored. These features are given to a number of regressors: SVR with Linear and RBF kernels, KNN with uniform weight, and DT. The Python Scikit-learn package is used to train the models. The train and test sets are described previously in Section 7.2.4.

Results

For regression, two metrics are used to measure the performance, RMSE and MAE, following the related work to compare the results obtained directly with theirs. Table 7.6 presents the results of depression severity regression for each feature individually. Interestingly, the best results are obtained using the HMSF, with 8.872 and 10.789 for MAE and RMSE, respectively. This indicates that head movements could play a significant role in estimating the severity level of depression.

The performance of feature fusion is measured, as described above. The fusion of the EBR with other features implies using type 3 of the SM with a different number of thresholds (i.e., 5, 50 and 500). The experimental results are shown in Table 7.7. It can be seen that the fusion of all the features achieves the best results with MAE and RMSE, which resulted from increasing the training set size from 50 video clips to 100 using both partitions: training and development. Notably, the best result is obtained when the EBR features are computed using five features and the number of consecutive frames is from at least two to a maximum of 5 frames. Alghowinem et al. (2013a) found that people with depression have a longer average duration of eye closure than healthy controls due to potential fatigue or the person is

Table 7.6 Results of depression regression on testing partition using each feature individually. The mean absolute error (MAE) and root mean square error (RMSE) are used to measure the performance ($linear-SVR^1, rbf-SVR^2, kNN^3, DT^4, KR^5$, i: is the number of consecutive frames, Dev: Development).

SM Type/Feature	Training set	No. of features	RMSE	MAE
EBR	Train	500 i=2:25	11.28^3	9.46
	Dev	5 i=2:9	14.20^3	12.21
	Train+Dev	50 i=2:28	12.68^3	9.69
HTR	Train	1	11.64^2	9.54
	Dev	1	11.57^3	9.77
	Train+Dev	1	11.57^5	9.50
HTR+HTSF	Train	13	11.65^2	9.54
	Dev	13	11.52^1	9.53
	Train+Dev	13	11.53^4	9.55
HMSF	Train	54	11.63^1	9.74
	Dev	54	11.14^3	9.47
	Train+Dev	54	**10.79^3**	**8.87**

trying to avoid eye contact. Also, the number of non-depressed people is larger than that of depressed people in the AVEC 2014 dataset. Considering all the factors mentioned above, when the length of the eye closure is small with almost five consecutive frames, it may help to distinguish non-depressed from depressed people. Alghowinem et al. (2013b) and Zhou et al. (2015) reported that head movement is a significant cue in detecting depression.

Comparing Performance with Related Work

In this chapter, the research focuses only on the visual-based approaches for depression regression. The performance obtained is compared with related work that also used the AVEC 2014 dataset. The visual features performance is comparable to related work that used advanced techniques (e.g., neural networks). Even though previous work employed advanced techniques, hand-crafted features are used in this current research because of the small size of the data. It can be seen from Table 7.8 that the approach in this thesis outperforms that of the previous work that used hand-crafted features (Jan et al., 2014; Kaya et al., 2014;

Table 7.7 Results of depression regression on the testing partition for features fusion and selection. The mean absolute error (MAE) and root mean square error (RMSE) are used to measure the performance ($linear-SVR^1, rbf-SVR^2, kNN^3, DT^4$, i: is the number of consecutive frames, Dev: Development).

Feature	Training set	No. of features	RMSE	MAE
EBR+HTR	Train	501 i=2:25	11.30^3	9.47
	Dev	6 i=2:9	13.82^3	11.50
	Train+Dev	51 i=2:10	11.74^3	9.50
EBR+HTR+HTSF	Train	513 i=2:14	12.56^2	9.75
	Dev	513 i=2:10	12.02^3	10.04
	Train+Dev	18 i=2:5	11.10^2	9.18
EBR+HMSF	Train	59 i=2:6	11.66^2	9.94
	Dev	59 i=2:9	10.79^3	8.97
	Train+Dev	554 i=2:21	12.29^2	9.93
EBR+HTR+HTSF+HMSF	Train	72 i=2:3	11.66^2	9.93
	Dev	117 i=2:3	11.49^3	9.55
	Train+Dev	72 i=2:5	**10.46^3**	**8.31**
Selected features (RFECV)	Train+Dev	45 i=2:5	**10.23^3**	**8.10**

Valstar et al., 2014). In addition, the proposed approach achieved better results than several methods that use audio and video data (Pérez Espinosa et al., 2014). The last four proposed models, which were built based on deep learning techniques, outperform other methods; in particular, Jan et al. (2017) achieved the best results among other methods in terms of RSME with 8.04, and Zhou et al. (2019) obtained the best results among other solutions in terms of MAE with 6.21.

Table 7.8 Depression regression results compared to other methods on AVEC2014 (Test Set) (PLS: partial least square, LinearReg: linear regression, DNN: deep neural networks).

Paper	Approach	Regressor	RMSE	MAE
Valstar et al. (2014) (Baseline)	Hand-crafted features	SVR	10.86	8.86
Jan et al. (2014)	Hand-crafted features	PLS+LR	10.50	8.44
Kaya et al. (2014)	Hand-crafted features	PCA+MPGI	10.27	8.20
The proposed approach	Hand-crafted features	KNN	10.23	8.10
Zhu et al. (2017)	DNN	DNN	9.55	7.47
Jan et al. (2017)	DNN	PLS+LR	**8.04**	6.68
Zhou et al. (2018)	DNN	DNN	8.43	6.37
Zhou et al. (2019)	DNN	DNN	8.39	**6.21**

A strong relationship between eye movements and depression recognition has been reported in the literature. When researchers used a frame that included only the eye region and fed this to convolutional neural networks, the best-obtained regression results were comparable to those obtained using other face regions (Zhou et al., 2019). Together these results provide important insights into the importance of these hand-crafted features and their ability to provide valuable information. However, the results are affected by the imbalanced training data, as shown in Figure 7.2. In other words, the training partition has a larger number of people with no depression than people with depression (mild, moderate, or severe depression levels).

The possible explanation for having less performance than deep learning techniques is that hand-crafted features such as the EBR and head movements require long videos to capture important information, whereas the AVEC 2014 dataset's video clips are short. There are, in fact, fewer frames in the training and development partitions than in the test partition (see Table 7.9). Finally, as stated at the beginning of this chapter, the main goal is to validate this research's approach on a public dataset and explore its feasibility, not to compete with state-of-the-art approaches.

Table 7.9 The number of frames in each set of the AVEC 2014 dataset.

Training	Development	Testing	Total
94,981	80,190	119,880	295,051

7.4 Conclusion

In this chapter, work to validate the proposed approach on a standardised depression dataset was conducted by investigating the use of hand-crafted features, such as the EBR with the MTs approach, the HTR, the HTSF and the HMSF. This investigation involved two tasks: classification and regression. For the classification task, the EBR achieved the highest performance in identifying depressed from non-depressed people, with 68% accuracy. The HTR, with or without HTSF, gave a chance-level performance, which is in agreement with the results obtained by Alghowinem et al. (2013b), who found that people with depression tend not to turn their heads to the left or right side, but downwards. The feature fusion performed better than only the HMSF. However, feature selection did not enhance the performance of the classifiers. Importantly, these findings seem to be consistent with those of previous studies, which found that these cues could be used as indicators of depression (Alghowinem et al., 2020; Gupta et al., 2014; Ringeval et al., 2019; Zhou et al., 2015)

In terms of the regression task, the HMSF gave the best results when each feature was tested individually. However, when the visual features were combined, they achieved the highest performance and using the feature selection improved the performance. The results indicate that using a combination of all these feature vectors is important to improve the regression performance due to the small data size. These results are consistent with those of Zhou et al. (2019), who found that motion in the eye region showed significant performance in estimating the severity of depression. That is why the eye blink is considered an important feature. It was observed from all the experiments that the performance of those visual features is comparable to related work that used advanced techniques. In addition, the results showed that using only visual-based approaches could achieve performance comparable to the state-of-the-art approaches in the challenge of AVEC 2014, which used both audio and video modalities. Issues with this dataset are its small size, short video clips and imbalanced training data set. Despite the relatively limited sample, this work offers valuable insights into the importance of these visual cues in healthcare applications.

Chapter 8

Conclusion

"There is hope, even when your brain tells you there isn't."
— John Green

This thesis has investigated the feasibility of using facial features, specifically eye blink rate (EBR), head turn rate (HTR), head turn statistical features (HTSF) and head movement statistical features (HMSF), for automatically detecting cognitive impairment (CI) using in-the-wild video data. This kind of data can pose several challenges, such as low resolution, poor illumination, participants' spontaneous behaviour during the session and noisy background. These challenges can lead to losing the participant's face, resulting in calculating false values for the facial landmarks and the head movements. This can become more challenging in the case of the appearance of more than one person in the camera view because the accompanying person can be closer to the camera than the participant him/herself, making the face detector detects only that person instead of the participant. This thesis, therefore, aims to develop a system for automatic CI detection. This system was conducted in three steps. The first step was to conduct a pre-processing operation for the videos that contained the appearance of people closer to the camera view than the participant. The second step was the feature extraction of facial cues. The third step was to feed these features to classifiers, producing a label for each participant.

In order to achieve this aim, this research was divided into three tasks. The first task focused on investigating the system for CI detection, as mentioned above, on a small dataset of people with CI. The second task focused on evaluating this research work on a larger dataset of people with CI and healthy controls (HC) with more variation in in-the-wild conditions. The third task involved validating the specific approach in this research on a public dataset to enable a comparison of the performance of this work with related work. Section 8.1 summarises how these tasks were investigated, along with the findings. Section 8.2 presents the potential directions of future work.

8.1 Summary of Thesis

Three research questions were investigated to achieve the tasks mentioned above:

1. What are the kinds of challenges and diversity that should be included in in-the-wild datasets to make them as representative as possible of real-world environments? (**Chapter 3**)

2. How can facial features be automatically detected in a robust way for in-the-wild data? (**Chapters 4 and 5**)

3. How useful are eye blink rate and head movement for CI detection? (**Chapters 4, 5, 6 and 7**)

First, Chapter 3 presented that the ‘in-the-wild data’ term should cover many of the edge cases of the participants in terms of demographic, look and behaviour, environmental conditions and device used (**RQ.1**). Then, it provided a review of different types of commonly used datasets for healthy individuals and people with health conditions, including their limitations. There was a lack of information about the challenges included in those datasets in terms of the definition of in-the-wild data, and most of them used professional cameras to record the data. Regarding the datasets for people with health conditions, most previous researchers evaluated their work on data recorded in a lab-controlled environment. This chapter also highlighted the importance of collecting in-the-wild datasets and the barriers to doing that.

Chapter 3 also introduced the datasets used in this research: IVA_{18}, IVA_{34}, and IVA_{52}, which was a combination of IVA_{18} and IVA_{34}. A comparison of the challenges of these datasets with those used in previous work was made. This comparison helped to show how these research datasets comprised in-the-wild data. It also demonstrated that these research in-the-wild datasets are the first kind to be used in both the computer vision and healthcare fields to automatically detect health conditions. This helped to provide insight into the significance of using such a dataset to develop a home-based application in the future.

Chapter 4 described the three steps mentioned above by starting to conduct a pre-processing operation on the video recordings of the IVA_{18} dataset. Then, two facial landmarks tracking techniques, Dlib and the OpenFace, were used to estimate the facial landmarks and then used the eyes’ landmarks to calculate the eye aspect ratio (EAR). The IVA_{18} dataset can be considered an in-the-wild dataset, which, as described in Chapter 3, could affect the EBR calculation because it depends on the EAR. Therefore, Chapter 4 presented a novel multiple thresholds (MTs) approach to calculate the EBR feature for the IVA_{18} dataset (**RQ.2**). This approach calculates multiple thresholds for blink detection, resulting in having multiple blinks for a specific range for each participant. This approach is more robust for in-the-wild data challenges. This approach was investigated using Dlib and OpenFace, to explore the performance on such an in-the-wild dataset. After the EBR was calculated, it was fed to classifiers. The results showed that Dlib facial landmarks gave better results than OpenFace, with an accuracy of 89% and 78%, respectively, for a three-way classification problem. The findings showed that EBR is a valuable indicator for differentiating neurodegenerative disorder (ND), mild cognitive impairment (MCI) and functional memory disorder (FMD) (**RQ.3**).

The MTs approach was then evaluated using a larger dataset for people with CI (IVA_{52}). First, the feasibility of using the MTs approach on the IVA_{52} data was investigated for classifying people with ND, MCI, FMD and HC in four-way, three-way and a combination of two-way problems. When the MTs approach was applied to the combined dataset, it showed an issue related to the classification of the newly added recordings (IVA_{34}) by classifying all of them incorrectly. The reason behind this was the variation in the recording environments and the devices used, resulting in having a lower mean and SD of the IVA_{34} participants than those in the IVA_{18} dataset. The MTs approach was, therefore, further developed to overcome these issues by calculating the threshold participant-dependent (PD) threshold (i.e., the thresholds for each participant were calculated based on their minimum and maximum values) instead of using the overall participants' maximum and minimum (**RQ.2 & RQ.3**). Then, different outlier setup approaches were used to deal with the extremely high values in the calculated EAR. The results showed improvements in the performance of the classification for the four-way, three-way and the combination of two-way problems.

Next, Chapter 6 described the first attempt to examine the feasibility of automatically detecting ND, MCI and FMD using head movement features, which was conducted on the IVA_{18} dataset (**RQ.3**). The features extracted were the HTR, the HTSF, the HMSF and the low-level features (LLFs). The performance was measured using each feature individually, the fusion of these head movement features with the EBR feature, and then when the feature selection was applied, and the LLFs. The results showed that some classification problems were considered challenging. For example, classifying ND vs. MCI and MCI vs. FMD using the HTR feature achieved an accuracy of only 67%, whereas classifying ND from FMD gave a good accuracy score of 92%. The use of the HMSF feature achieved high performance in differentiating MCI from FMD, with an accuracy of 83%. The feature fusion helped to improve the classification results of a three-way problem with an accuracy of 78%.

On the other hand, the LLFs did not improve the classification results for three-way and two-way problems compared with other features. The findings suggested a relationship between increasing HTR and the progression of MCI and dementia to Alzheimer's disease. It was also found that the presence of an accompanying person plays a significant role in increasing the HTR. The findings were very encouraging for data recorded in the wild compared with data recorded in a lab-controlled condition.

The feasibility of the head movement features, HTR, HTSF and HMSF, for the automatic CI detection on IVA_{52} was explored (**RQ.3**). The experiment involved classification problems that were a combination of four-way, three-way and two-way problems. The results showed that the features that achieved the highest accuracy in the IVA_{52} dataset were the same features that obtained the highest performance in the IVA_{18}. It was clear that the HTR feature and its derivative features (HTSF) gave the highest performance in classifying three-way and ND from FMD with an accuracy of 59% and 92%, respectively. It also showed that using the HMSF features gave the highest performance in classifying ND from MCI and MCI from FMD, with an accuracy of 62% and 90%, respectively. This indicates that HTR+HTSF is a very valuable cue for differentiating CI from other health conditions. Moreover, the HMSF could help to differentiate health conditions with similar symptoms. Finally, the classification problems showed promising results for the other two-way problems. However, it showed that classifying FMD from HC was a very challenging case with a 67% accuracy in comparison with other classification problems when ND was classified from HC with 83% accuracy and MCI from HC with a 74% accuracy. The findings provided an insight into the difficulty of classifying these four health conditions from each other compared to previous work that had investigated only one two-way classification problem: classifying CI (regardless of CI type and including MCI and ND as one group) from HC. They also provided insight into the value of such facial features for detecting health conditions using in-the-wild data.

The MTs approach used in Chapter 4 was validated in Chapter 7 using a public depression dataset (AVEC 2014). The aim of this was to examine the applicability of the findings to another condition that shares similar symptoms to those of CI (**RQ.3**). This investigation used the EBR with the MTs approach and head movement features, HTR, HTSF, and HMSF, which was described in Chapter 6. The validation involved two tasks, classification and regression. In terms of classification, the results showed that the EBR feature classified the depressed group (DG) from the non-depressed group (NDG), with an accuracy of 68%. However, the HTR features achieved an accuracy of only a chance level because people with depression do not usually turn their heads to the left or right side, only downward, to avoid eye contact with the interviewer (Fossi et al., 1984). It was found that the feature selection did not show any enhancement in the results compared with the individual feature results. For the regression task, only the HMSF achieved the highest results when individual features measured the performance. The feature fusion did not improve the performance until feature selection was applied. The findings showed that employing all features contributed significantly to improving the performance of the regression. To conclude, the performance obtained using these hand-crafted features was comparable to that reported in previous studies that had used advanced techniques (i.e., neural networks). It was also shown that facial features were comparable to the state-of-art approaches that employed both modalities (audio and video).

In conclusion, this thesis has far-reaching benefits for researchers and clinicians. It can improve their understanding of these health conditions, facilitate novel discoveries, and improve detection accuracy. This can result in the ability to differentiate between different health conditions that share similar symptoms and conduct more deeply focused and efficient research outcomes, which lead to providing patients with suitable care and treatment.

8.2 Scope for Future Work

This thesis is the first to investigate different health conditions (ND, MCI, FMD and HC) using facial features from video recordings. Thus, this can be a start for researchers to conduct more deep investigations because there are more challenges and areas of improvement that need to be tackled and more aspects to be discovered in the detection of CI to enhance the overall performance of the system. Several possible future work areas for this research are suggested below.

Increasing the size of the dataset

The number of participants with health conditions in the IVA_{52} dataset needs to be increased by recruiting and recording more participants with ND, MCI and FMD from both genders with matched ages. Having more data will give an opportunity to conduct more investigations based on health condition, gender and the presence of a partner, which may help to improve the classification results. As a result of various issues with video recordings in the IVA_{52} dataset that led to some of them having to be excluded, data should be recorded with some instructions for the participants to allow for the accurate detection of eye landmarks in in-the-wild data. The aim of these instructions is to reduce the false values of the eye landmarks detection that result in high values for the EAR because the face detector loses track of the facial landmarks. The instructions should detail how the participants should sit and the room's lighting or the use of daylight. For instance, they should sit in front of the laptop or smartphone, and their faces should appear very clearly in the camera, and they should sit in a room with good light, which should not be behind the participant or directly on top of the participant's head. These cases result in a dark area for the face region, which makes eye detection very difficult and can result in false values for eye landmarks detection.

Annotating and analysing the dataset

According to Taati et al. (2019), state-of-the-art approaches for face and facial landmark detection fail or have lower performance when applied to datasets for elderly people with CI. Further work can be carried out by allocating ample time for annotating the eye status (i.e., closed, open, partially open), head movement and every existing condition in terms of the participant's look and behaviour and environment in each video frame for each participant. Since the dataset is very challenging, conducting deep analyses of such datasets will help to 1) understand why those approaches had such performance and 2) paint a full picture of the reason for any performance obtained using any approach for capturing particular facial features.

Investigating deep learning techniques

After collecting more data, another approach could be used for exploring the classification performance of different deep learning models. This investigation could be conducted by training a neural network model on those facial features. In addition, they could be trained on video frames by evaluating the performance using the whole frame cropped to include only the participant's face and in which the frame is divided into three parts: upper (forehead and eyebrows), middle (the eyes region) and lower (the mouth). Dividing the frames into parts this way could help to determine which part of the face plays a crucial role in the classification performance.

In this thesis, the initial preprocessing phase for the IVA dataset is conducted manually. However, there is potential to automate this step by employing deep learning techniques to extract facial features from each frame of the participant's face. The next step involves calculating the cosine similarity among these facial features for each frame of the same participant. A feature vector representing a neutral and clear facial expression is established as a reference point for comparison with other feature vectors from different frames. The resulting values will indicate the dissimilarity or similarity between the facial features, with a range of -1 to 1, where 1 signifies identical vectors, and -1 signifies opposite vectors. This approach may be useful in finding the frames where the patient's face is not in the camera view and the frames that represent only the patient.

Investigating more facial features

In this thesis, EBR and head movement features were examined. There are more derivative features related to each one of them. For example, more features can be extracted for each eye blink detected, such as the blink's duration and frequency. The blink duration can be measured by the time interval from the eyelids closing and reopening. Blink frequency determines how many times a person can blink within a particular time frame. It is calculated using the time intervals between genuine detected blinks. These features can be calculated using the MTs approach. More features can be extracted regarding head movements, such as calculating the minimum, maximum, range and average duration of looking: right, left, up and down, and tilting clockwise and anti-clockwise for all head movements. Also, the total number of changes in looking direction for pitch, yaw, roll and all movements can be investigated. Exploring more facial features, eye gaze, facial expressions and facial action units, could all help to enhance the classification results. It would also be beneficial for the system's performance if facial landmarks are employed to extract features, such as facial expressions and statistical features.

Investigating multi-modal system

The applicability of building a multi-modal system to enhance the performance of the classification could be investigated. A multi-modal system could be built in different ways by investigating speech-based features and then combining them with facial features, which may result in a language-independent multi-modal system. It could also be built by investigating the combination of language, speech, audio and facial features as one multi-modal system.

Developing an application for CI detection

Work has already been carried out to develop a system for detecting early signs of dementia and FMD (CognoSpeak [1]) (Brewer et al., 2021; Mirheidari et al., 2022; O'Malley et al., 2021). It uses speech technology and machine learning to extract features of a person's speech and use them to detect CI. The promising results of the current research obtained in this thesis could be integrated with the CognoSpeak system. This would help to achieve the purpose of this research, which is to have a tool which is low in cost, effort and time, non-invasive, accessible by other people, and capable of being used in a patient's home.

[1] www.cognospeak.com

References

Abate, A. F., Bisogni, C., Castiglione, A., and Nappi, M. (2022). Head pose estimation: An extensive survey on recent techniques and applications. *Pattern Recognition*, 127:108591.

Ahmad, A., Owais, K., Siddiqui, M., Mamun, K., Rao, F., and Yousufzai, A. W. (2013). Dementia in pakistan: national guidelines for clinicians. *Pakistan Journal of Neurological Sciences (PJNS)*, 8(3):17–27.

Al-Rahayfeh, A. and Faezipour, M. (2013a). Eye Tracking and Head Movement Detection: A State-of-Art Survey. *IEEE Journal of Translational Engineering in Health and Medicine*, 1(November):2100212–2100212.

Al-Rahayfeh, A. and Faezipour, M. (2013b). Eye tracking and head movement detection: A state-of-art survey. *IEEE journal of translational engineering in health and medicine*, 1:2100212–2100212.

Alberdi, A., Aztiria, A., and Basarab, A. (2016). On the early diagnosis of alzheimer's disease from multimodal signals: A survey. *Artificial intelligence in medicine*, 71:1–29.

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., et al. (2011). The diagnosis of mild cognitive impairment due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia*, 7(3):270–279.

Alghowinem, S., Goecke, R., Cohn, J. F., Wagner, M., Parker, G., and Breakspear, M. (2015). Cross-cultural detection of depression from nonverbal behaviour. In *2015 11th IEEE International conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–8. IEEE.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G., et al. (2012). From joyous to clinically depressed: Mood detection using spontaneous speech. In *FLAIRS Conference*, volume 19. Citeseer.

Alghowinem, S., Goecke, R., Wagner, M., Parker, G., and Breakspear, M. (2013a). Eye movement analysis for depression detection. In *2013 IEEE International Conference on Image Processing*, pages 4220–4224. IEEE.

Alghowinem, S., Goecke, R., Wagner, M., Parker, G., and Breakspear, M. (2013b). Head pose and movement analysis as an indicator of depression.

Alghowinem, S. M., Gedeon, T., Goecke, R., Cohn, J., and Parker, G. (2020). Interpretation of depression detection models via feature selection methods. *IEEE Transactions on Affective Computing*.

Alhakeem, Z. M., Ali, R. S., and Abd-Alhameed, R. A. (2020). Wheelchair free hands navigation using robust dwt_ar features extraction method with muscle brain signals. *IEEE Access*, 8:64266–64277.

Allan, C. L., Behrman, S., Ebmeier, K. P., and Valkanova, V. (2017). Diagnosing early cognitive decline—when, how and for whom? *Maturitas*, 96:103–108.

Alsaeedi, N. and Wloka, D. (2019). Real-time eyeblink detector and eye state classifier for virtual reality (vr) headsets (head-mounted displays, hmds). *Sensors*, 19(5):1121.

Alz.org (2019). Types of dementia.

Anas, E. R., Henriquez, P., and Matuszewski, B. J. (2017). Online eye status detection in the wild with convolutional neural networks. In *International Conference on Computer Vision Theory and Applications*, volume 7, pages 88–95. SciTePress.

Anderson, T. J. and MacAskill, M. R. (2013). Eye movements in patients with neurodegenerative disorders. *Nature Reviews Neurology*, 9(2):74–85.

Appel, T., Santini, T., and Kasneci, E. (2016). Brightness-and motion-based blink detection for head-mounted eye trackers. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 1726–1735. ACM.

Arcoverde Neto, E. N., Duarte, R. M., Barreto, R. M., Magalhães, J. P., Bastos, C., Ren, T. I., and Cavalcanti, G. D. (2014). Enhanced real-time head pose estimation system for mobile device. *Integrated Computer-Aided Engineering*, 21(3):281–293.

Argilés, M., Cardona, G., Pérez-Cabré, E., and Rodríguez, M. (2015). Blink rate and incomplete blinks in six different controlled hard-copy and electronic reading conditions. *Investigative ophthalmology & visual science*, 56(11):6679–6685.

Arnáiz, E. and Almkvist, O. (2003). Neuropsychological features of mild cognitive impairment and preclinical alzheimer's disease. *Acta Neurologica Scandinavica*, 107:34–41.

Asgarian, A., Zhao, S., Ashraf, A. B., Browne, M. E., Prkachin, K. M., Mihailidis, A., Hadjistavropoulos, T., and Taati, B. (2019). Limitations and biases in facial landmark detection d an empirical study on older adults with dementia. In *CVPR Workshops*, pages 28–36.

Asplund, K., Norberg, A., Adolfsson, R., and Waxman, H. M. (1991). Facial expressions in severely demented patients—a stimulus–response study of four patients with dementia of the alzheimer type. *International Journal of Geriatric Psychiatry*, 6(8):599–606.

Association, A. P. et al. (2013). Diagnostic and statistical manual of mental disorders. *BMC Med*, 17:133–137.

Azim, T., Jaffar, M. A., and Mirza, A. M. (2014). Fully automated real time fatigue detection of drivers through fuzzy expert systems. *Applied Soft Computing*, 18:25–38.

Babacan-Yildiz, G., Isik, A. T., Ur, E., Aydemir, E., Ertas, C., Cebi, M., Soysal, P., Gursoy, E., Kolukisa, M., Kocaman, G., et al. (2013). Cost: Cognitive state test, a brief screening battery for alzheimer disease in illiterate and literate patients. *International Psychogeriatrics*, 25(3):403–412.

Bacivarov, I., Ionita, M., and Corcoran, P. (2008). Statistical models of appearance for eye tracking and eye-blink detection and measurement. *IEEE transactions on consumer electronics*, 54(3):1312–1320.

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2012). 3d constrained local model for rigid and non-rigid facial tracking. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2610–2617. IEEE.

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

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE.

Barra, P., Barra, S., Bisogni, C., De Marsico, M., and Nappi, M. (2020). Web-shaped model for head pose estimation: An approach for best exemplar selection. *IEEE Transactions on Image Processing*, 29:5457–5468.

Barral, O., Jang, H., Newton-Mason, S., Shajan, S., Soroski, T., Carenini, G., Conati, C., and Field, T. (2020). Non-invasive classification of Alzheimer's disease using eye tracking and language. In *Machine Learning for Healthcare Conference*, pages 813–841. PMLR.

Bartlett, M. S., Littlewort, G., Frank, M. G., Lainscsek, C., Fasel, I. R., Movellan, J. R., et al. (2006). Automatic recognition of facial actions in spontaneous expressions. *J. Multim.*, 1(6):22–35.

Bayles, K. A., Kaszniak, A. W., and Tomoeda, C. K. (1987). *Communication and cognition in normal aging and dementia.* College-Hill Press/Little, Brown & Co.

Beck, A. T., Steer, R. A., and Brown, G. (1996). Beck depression inventory–ii. *Psychological Assessment.*

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., and Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940.

Beltrán, J., García-Vázquez, M. S., Benois-Pineau, J., Gutierrez-Robledo, L. M., and Dartigues, J.-F. (2018). Computational techniques for eye movements analysis towards supporting early diagnosis of alzheimer's disease: A review. *Computational and mathematical methods in medicine*, 2018.

Bentivoglio, A. R., Bressman, S. B., Cassetta, E., Carretta, D., Tonali, P., and Albanese, A. (1997). Analysis of blink rate patterns in normal subjects. *Movement disorders*, 12(6):1028–1034.

Beville, B. P. K. (2012). Dementia Simulation Methods and systems for simulation of cognitive decline. pages 1–9.

Bhowmick, S. and Mustafa, H. A. (2021). A framework for eye-based human machine interface. In *2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pages 1–6. IEEE.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Boholm, M. and Allwood, J. (2010). Repeated head movements, their function and relation to speech. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 6–10. Citeseer.

Bonello, M. and Larner, A. (2016). Applause sign: screening utility for dementia and cognitive impairment. *Postgraduate medicine*, 128(2):250–253.

Bornstein, N. M., Brainin, M., Guekht, A., Skoog, I., and Korczyn, A. D. (2014). Diabetes and the brain: issues and unmet needs. *Neurological Sciences*, 35(7):995–1001.

Boucart, M., Bubbico, G., Szaffarczyk, S., and Pasquier, F. (2014). Animal spotting in alzheimer's disease: an eye tracking study of object categorization. *Journal of Alzheimer's Disease*, 39(1):181–189.

Bouchard, R. and Rossor, M. (2007). Typical clinical features. Informa Healthcare.

Braak, H. and Braak, E. (1997). Frequency of stages of alzheimer-related lesions in different age categories. *Neurobiology of aging*, 18(4):351–357.

Braak, H., Rüb, U., Steur, E. J., Del Tredici, K., and De Vos, R. (2005). Cognitive status correlates with neuropathologic stage in parkinson disease. *Neurology*, 64(8):1404–1410.

Brewer, E., Mirheidari, B., O'Malley, R., Reuber, M., Christensen, H., and Blackburn, D. J. (2021). Characterising spoken interactions of healthy ageing adults with cognospeak, a web-based cognitive assessment tool. *Alzheimer's & Dementia*, 17:e052913.

Brown, L. M. and Tian, Y.-L. (2002). Comparative study of coarse head pose estimation. In *Workshop on Motion and Video Computing, 2002. Proceedings.*, pages 125–130. IEEE.

Burianova, H., McIntosh, A. R., and Grady, C. L. (2010). A common functional brain network for autobiographical, episodic, and semantic memory retrieval. *Neuroimage*, 49(1):865–874.

Burl, M. C. and Perona, P. (1996). Recognition of planar object classes. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 223–230. IEEE.

Busso, C. and Jain, J. (2012). Advances in multimodal tracking of driver distraction. In *Digital Signal Processing for In-Vehicle Systems and Safety*, pages 253–270. Springer.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE.

Carter, M. J. (2014). Diagnostic and statistical manual of mental disorders. *Therapeutic recreation journal*, 48(3):275.

Celestino, J., Marques, M., Nascimento, J. C., and Costeira, J. P. (2023). 2d image head pose estimation via latent space regression under occlusion settings. *Pattern Recognition*, 137:109288.

Cermakova, P., Ding, J., Meirelles, O., Reis, J., Religa, D., Schreiner, P. J., Jacobs Jr, D. R., Bryan, R. N., and Launer, L. J. (2020). Carotid intima–media thickness and markers of brain health in a biracial middle-aged cohort: Cardia brain mri sub-study. *The Journals of Gerontology: Series A*, 75(2):380–386.

Cermakova, P., Muller, M., Armstrong, A. C., Religa, D., Bryan, R. N., Lima, J. A., and Launer, L. J. (2017a). Subclinical cardiac dysfunction and brain health in midlife: Cardia (coronary artery risk development in young adults) brain magnetic resonance imaging substudy. *Journal of the American Heart Association*, 6(12):e006750.

Cermakova, P., Nelson, M., Secnik, J., Garcia-Ptacek, S., Johnell, K., Fastbom, J., Kilander, L., Winblad, B., Eriksdotter, M., and Religa, D. (2017b). Living alone with alzheimer's disease: data from svedem, the swedish dementia registry. *Journal of Alzheimer's Disease*, 58(4):1265–1272.

Chau, M. and Betke, M. (2005). Real time eye tracking and blink detection with usb cameras. Technical report, Boston University Computer Science Department.

Chen, J. and Amayeh, G. (2019). Detailed eye shape model for robust biometric applications. US Patent App. 15/693,975.

Chen, S., Zhang, Y., Yin, B., and Wang, B. (2021). Trfh: towards real-time face detection and head pose estimation. *Pattern Analysis and Applications*, 24:1745–1755.

Chen, X., Lu, Y., Cao, B., Lin, D., and Ahmad, I. (2023). Lightweight head pose estimation without keypoints based on multi-scale lightweight neural network. *The Visual Computer*, pages 1–15.

Chermahini, S. and Hommel, B. (2012). Kreativniji kroz pozitivno raspoloženje. nitko. *Ispred. Pjevušiti. Neurosci*, 6(319.10):3389.

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

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

Cortacero, K., Fischer, T., and Demiris, Y. (2019). Rt-bene: a dataset and baselines for real-time blink estimation in natural environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

Costanza, A., D'Orta, I., Perroud, N., Burkhardt, S., Malafosse, A., Mangin, P., and La Harpe, R. (2014). Neurobiology of suicide: do biomarkers exist? *International journal of legal medicine*, 128:73–82.

Coubard, O. A. (2016). What do we know about eye movements in alzheimer's disease? the past 37 years and future directions.

Crawford, T. J., Higham, S., Renvoize, T., Patel, J., Dale, M., Suriya, A., and Tetley, S. (2005). Inhibitory control of saccadic eye movements and cognitive impairment in alzheimer's disease. *Biological psychiatry*, 57(9):1052–1060.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.

De Jong, P. J. and Merckelbach, H. (1990). Eyeblink frequency, rehearsal activity, and sympathetic arousal. *International Journal of Neuroscience*, 51(1-2):89–94.

de la Cruz, G., Lira, M., Luaces, O., and Remeseiro, B. (2022). Eye-lrcn: A long-term recurrent convolutional network for eye blink completeness detection. *IEEE Transactions on Neural Networks and Learning Systems*.

de Lima Medeiros, P. A., da Silva, G. V. S., dos Santos Fernandes, F. R., Sánchez-Gendriz, I., Lins, H. W. C., da Silva Barros, D. M., Nagem, D. A. P., and de Medeiros Valentim, R. A. (2022). Efficient machine learning approach for volunteer eye-blink detection in real-time using webcam. *Expert Systems with Applications*, 188:116073.

De Padova, V., Barbato, G., Conte, F., and Ficca, G. (2009). Diurnal variation of spontaneous eye blink rate in the elderly and its relationships with sleepiness and arousal. *Neuroscience letters*, 463(1):40–43.

Delgado-García, J. M., Gruart, A., and Múnera, A. (2002). Neural organization of eyelid responses. *Movement disorders*, 17(S2):S33–S36.

Delgado-García, J. M., Gruart, A., and Trigo, J. A. (2003). Physiology of the eyelid motor system. *Annals of the New York Academy of Sciences*, 1004(1):1–9.

DeMenthon, D. F. and Davis, L. S. (1995). Model-based object pose in 25 lines of code. *International journal of computer vision*, 15(1):123–141.

Dewi, C., Chen, R.-C., Jiang, X., and Yu, H. (2022). Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks. *PeerJ Computer Science*, 8:e943.

Dibeklioğlu, H., Hammal, Z., and Cohn, J. F. (2017). Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE journal of biomedical and health informatics*, 22(2):525–536.

Dibeklioğlu, H., Hammal, Z., Yang, Y., and Cohn, J. F. (2015). Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 307–310.

Divjak, M. and Bischof, H. (2009). Eye blink based fatigue detection for prevention of computer vision syndrome. In *MVA*, pages 350–353.

Dong, W. and Wu, X. (2005). Fatigue detection based on the distance of eyelid. In *Proceedings of 2005 IEEE International Workshop on VLSI Design and Video Technology, 2005.*, pages 365–368. IEEE.

Draucker, C. B. (2005). Processes of mental health service use by adolescents with depression. *Journal of Nursing Scholarship*, 37(2):155–162.

Drutarovsky, T. and Fogelton, A. (2014). Eye blink detection using variance of motion vectors. In *European Conference on Computer Vision*, pages 436–448. Springer.

Duda, R. O., Hart, P. E., et al. (2006). *Pattern classification.* John Wiley & Sons.

Durães, J., Tábuas-Pereira, M., Araújo, R., Duro, D., Baldeiras, I., Santiago, B., and Santana, I. (2018). The head turning sign in dementia and mild cognitive impairment: its relationship to cognition, behavior, and cerebrospinal fluid biomarkers. *Dementia and Geriatric Cognitive Disorders*, 46(1-2):42–49.

Ebert, D., Albert, R., Hammon, G., Strasser, B., May, A., and Merz, A. (1996). Eye-blink rates and depression: Is the antidepressant effect of sleep deprivation mediated by the dopamine system? *Neuropsychopharmacology*, 15(4):332–339.

Eddine, B. D., Dos Santos, F. N., Boulebtateche, B., and Bensaoula, S. (2018). Eyelsd a robust approach for eye localization and state detection. *Journal of Signal Processing Systems*, 90(1):99–125.

Ekman, P., Friesen, W. V., and Hager, J. C. (2002). The facial action coding system: The manual on cd-rom & investigator's guide. *Salt Lake City, UT: Research Nexus.*

Ellgring, H. (2007). *Non-verbal communication in depression.* Cambridge University Press.

Elsawy, B. and Higgins, K. E. (2011). The geriatric assessment. *Am Fam Physician*, 83(1):48–56.

Endo, T., Ukita, N., Tanaka, H., Hagita, N., Nakamura, S., Adachi, H., Ikeda, M., Kazui, H., and Kudo, T. (2017). Initial response time measurement in eye movement for dementia screening test. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 262–265. IEEE.

Fahad, M. S., Ranjan, A., Yadav, J., and Deepak, A. (2021). A survey of speech emotion recognition in natural environment. *Digital signal processing*, 110:102951.

Fanelli, G., Dantone, M., Gall, J., Fossati, A., and Van Gool, L. (2013). Random forests for real time 3d face analysis. *International journal of computer vision*, 101(3):437–458.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer.

Fei, Z., Yang, E., Li, D. D.-U., Butler, S., Ijomah, W., and Zhou, H. (2019). A survey on computer vision techniques for detecting facial features towards the early diagnosis of mild cognitive impairment in the elderly. *Systems Science & Control Engineering*, 7(1):252–263.

Firintepe, A., Selim, M., Pagani, A., and Stricker, D. (2020). The more, the merrier? a study on in-car ir-based head pose estimation. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1060–1065. IEEE.

Fogelton, A. and Benesova, W. (2016). Eye blink detection based on motion vectors analysis. *Computer Vision and Image Understanding*, 148:23–33.

Fogelton, A. and Benesova, W. (2018). Eye blink completeness detection. *Computer Vision and Image Understanding*, 176:78–85.

Foggin, E. (2018). When to suspect dementia. *InnovAiT*, 11(5):241–248.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). “mini-mental state”: a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198.

Fossi, L., Faravelli, C., and Paoli, M. (1984). The ethological approach to the assessment of depressive disorders. *Journal of Nervous and Mental Disease*.

Fraser, K. C., Lundholm Fors, K., Eckerström, M., Öhman, F., and Kokkinakis, D. (2019). Predicting mci status from multimodal language data using cascaded classifiers. *Frontiers in aging neuroscience*, page 205.

Fukui, T., Yamazaki, T., and Kinno, R. (2011). Can the ‘head-turning sign’be a clinical marker of alzheimer’s disease. *Dementia and geriatric cognitive disorders extra*, 1(1):310–317.

Gan, C. L. (2020). Prognostics and health management of electronics: Fundamentals, machine learning, and the internet of things.

Gao, J., Yang, Y., Lin, P., and Park, D. S. (2018). Computer vision in healthcare applications.

Gawande, R. and Badotra, S. (2022). Deep-learning approach for efficient eye-blink detection with hybrid optimization concept. *International Journal of Advanced Computer Science and Applications*, 13(6).

Ghadiri-Sani, M. and Larner, A. (2013). Head turning sign for diagnosis of dementia and mild cognitive impairment: a revalidation. *J Neurol Neurosurg Psychiatry*, 84(11):e2–e2.

Ghiass, R. S., Arandjelović, O., and Laurendeau, D. (2015). Highly accurate and fully automatic head pose estimation from a low quality consumer-level rgb-d sensor. In *Proceedings of the 2nd workshop on computational models of social interactions: Human-Computer-Media communication*, pages 25–34.

Gifford, D. R. and Cummings, J. L. (1999). Evaluating dementia screening tests: methodologic standards to rate their performance.

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S. M., Hammal, Z., and Rosenwald, D. P. (2014). Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647.

Gogate, M., Dashtipour, K., Adeel, A., and Hussain, A. (2020). Cochleanet: A robust language-independent audio-visual model for real-time speech enhancement. *Information Fusion*, 63:273–285.

Goschke, T. and Bolte, A. (2014). Emotional modulation of control dilemmas: The role of positive affect, reward, and dopamine in cognitive stability and flexibility. *Neuropsychologia*, 62:403–423.

Gou, C. and Ji, Q. (2020). Coupled cascade regression from real and synthesized faces for simultaneous landmark detection and head pose estimation. *Journal of Electronic Imaging*, 29(2):023028–023028.

Gou, C., Zhou, Y., Xiao, Y., Wang, X., and Yu, H. (2022). Cascade learning for driver facial monitoring. *IEEE Transactions on Intelligent Vehicles*.

Gourier, N., Hall, D., and Crowley, J. L. (2004). Estimating face orientation from robust detection of salient facial structures. In *FG Net workshop on visual observation of deictic gestures*, volume 6, page 7. Citeseer.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., et al. (2014). The distress analysis interview corpus of human and computer interviews. Technical report, University of Southern California Los Angeles.

Grauman, K., Betke, M., Gips, J., and Bradski, G. R. (2001). Communication via eye blinks-detection and duration analysis in real time. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE.

Grauman, K., Betke, M., Lombardi, J., Gips, J., and Bradski, G. R. (2003). Communication via eye blinks and eyebrow raises: Video-based human-computer interfaces. *Universal Access in the Information Society*, 2(4):359–373.

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2008). Multi-pie. proceedings of the eighth. In *IEEE International Conference on Automatic Face and Gesture Recognition*.

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-pie. *Image and vision computing*, 28(5):807–813.

Gulliver, A., Griffiths, K. M., and Christensen, H. (2010). Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. *BMC psychiatry*, 10(1):1–9.

Gupta, R., Malandrakis, N., Xiao, B., Guha, T., Van Segbroeck, M., Black, M., Potamianos, A., and Narayanan, S. (2014). Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pages 33–40.

Gupta, S., Fahad, M. S., and Deepak, A. (2020). Pitch-synchronous single frequency filtering spectrogram for speech emotion recognition. *Multimedia Tools and Applications*, 79:23347–23365.

Gupta, V. (2018). Face detection–opencv, dlib and deep learning (c++/python). *Learn OpenCV*.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422.

Hadar, U., Steiner, T. J., Grant, E. C., and Rose, F. C. (1984). The timing of shifts of head postures during conservation. *Human Movement Science*, 3(3):237–245.

Han, Y.-J., Kim, W., and Park, J.-S. (2018). Efficient eye-blinking detection on smartphones: A hybrid approach based on deep learning. *Mobile Information Systems*, 2018.

Hansen, D. W. and Ji, Q. (2009). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500.

He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., Guo, C., Wang, H., Ding, S., Wang, Z., et al. (2022). Deep learning for depression recognition with audiovisual cues: A review. *Information Fusion*, 80:56–86.

Heilman, K. M. and Nadeau, S. E. (2022). Emotional and neuropsychiatric disorders associated with alzheimer's disease. *Neurotherapeutics*, 19(1):99–116.

Heishman, R. and Duric, Z. (2007). Using image flow to detect eye blinks in color videos. In *2007 IEEE Workshop on Applications of Computer Vision (WACV'07)*, pages 52–52. IEEE.

Hesch, J. A. and Roumeliotis, S. I. (2011). A direct least-squares (dls) method for pnp. In *2011 International Conference on Computer Vision*, pages 383–390. IEEE.

Heylen, D. (2006). Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, 3(03):241–267.

Holland, A. A. and Larner, A. (2013). Effects of gender on two clinical signs (attended alone and head turning) of use in the diagnosis of cognitive complaints. *Journal of the Neurological Sciences*, 333:e295–e296.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Hoque, M. and Picard, R. W. (2011). Acted vs. natural frustration and delight: Many people smile in natural frustration. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 354–359. IEEE.

Horackova, K., Kopecek, M., Mach, V., Kagstrom, A., Aarsland, D., Motlova, L. B., and Cermakova, P. (2019). Prevalence of late-life depression and gap in mental health service use across european regions. *European Psychiatry*, 57:19–25.

Hsu, H.-W., Wu, T.-Y., Wan, S., Wong, W. H., and Lee, C.-Y. (2018). Quatnet: Quaternion-based head pose estimation with multiregression loss. *IEEE Transactions on Multimedia*, 21(4):1035–1046.

Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*.

Huang, K.-Y., Wu, C.-H., Kuo, Y.-T., and Jang, F.-L. (2016). Unipolar depression vs. bipolar disorder: An elicitation-based approach to short-term detection of mood disorder. *depression*, 10:12.

Jan, A., Meng, H., Gaus, Y. F. A., Zhang, F., and Turabzadeh, S. (2014). Automatic depression scale prediction using facial expression dynamics and regression. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 73–80.

Jan, A., Meng, H., Gaus, Y. F. B. A., and Zhang, F. (2017). Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Transactions on Cognitive and Developmental Systems*, 10(3):668–680.

Jayalakshmi, T. and Santhakumaran, A. (2011). Statistical normalization and back propagation for classification. *International Journal of Computer Theory and Engineering*, 3(1):1793–8201.

Jebara, T. S. (1995). 3d pose estimation and normalization for face recognition. *Centre for Intelligent Machines, McGill University*.

JH Balsters, M., J Krahmer, E., GJ Swerts, M., and JJM Vingerhoets, A. (2012). Verbal and nonverbal correlates for depression: a review. *Current Psychiatry Reviews*, 8(3):227–234.

Jiang, Z., Luskus, M., Seyedi, S., Griner, E. L., Rad, A. B., Clifford, G. D., Boazak, M., and Cotes, R. O. (2022). Utilizing computer vision for facial behavior analysis in schizophrenia studies: A systematic review. *PloS one*, 17(4):e0266828.

Jonell, P., Moëll, B., Håkansson, K., Henter, G. E., Kucherenko, T., Mikheeva, O., Hagman, G., Holleman, J., Kivipelto, M., Kjellström, H., et al. (2021). Multimodal capture of patient behaviour for improved detection of early dementia: clinical feasibility and preliminary results. *Frontiers in Computer Science*, 3:642633.

Jongkees, B. J. and Colzato, L. S. (2016). Spontaneous eye blink rate as predictor of dopamine-related cognitive function—a review. *Neuroscience & Biobehavioral Reviews*, 71:58–82.

Kamanga, I. A. and Lyimo, J. M. (2022). Anti-spoofing detection based on eyeblink liveness testing for iris recognition.

Kavé, G. and Goral, M. (2018). Word retrieval in connected speech in alzheimer's disease: a review with meta-analyses. *Aphasiology*, 32(1):4–26.

Kaya, H., Çilli, F., and Salah, A. A. (2014). Ensemble cca for continuous emotion prediction. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 19–26.

Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874.

Kemppainen, N., Laine, M., Laakso, M., Kaasinen, V., Någren, K., Vahlberg, T., Kurki, T., and Rinne, J. (2003). Hippocampal dopamine d2 receptors correlate with memory functions in alzheimer's disease. *European journal of neuroscience*, 18(1):149–154.

Khan, K., Attique, M., Syed, I., Sarwar, G., Irfan, M. A., and Khan, R. U. (2019). A unified framework for head pose, age and gender classification through end-to-end face segmentation. *Entropy*, 21(7):647.

Khan, K., Khan, R. U., Leonardi, R., Migliorati, P., and Benini, S. (2021). Head pose estimation: A survey of the last ten years. *Signal Processing: Image Communication*, 99:116479.

Kim, K., Baltrusaitis, T., Zadeh, A., Morency, L.-P., and Medioni, G. (2016). Holistically constrained local model: Going beyond frontal poses for facial landmark detection. Technical report, University of Southern California, Institute for Robotics and Intelligent ....

Kim, K. W., Hong, H. G., Nam, G. P., and Park, K. R. (2017). A study of deep cnn-based classification of open and closed eyes using a visible light camera sensor. *Sensors*, 17(7):1534.

Kim, M., Kumar, S., Pavlovic, V., and Rowley, H. (2008). Face tracking and recognition with visual constraints in real-world videos. In *2008 IEEE Conference on computer vision and pattern recognition*, pages 1–8. IEEE.

Kim, S. and Kang, Y. (2017). Development and Validation of the Way-Finding Ability Scale for Middle-Aged and Older Adults. *Dementia and Neurocognitive Disorders*, 16(4):95.

Kimball, A. L. (1917). *A college text-book of physics*. H. Holt.

King, D. C. and Michels, K. M. (1957). Muscular tension and the human blink rate. *Journal of experimental psychology*, 53(2):113.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758.

Kirk, M. and Berntsen, D. (2018). The life span distribution of autobiographical memory in alzheimer's disease. *Neuropsychology*, 32(8):906.

Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., and Jain, A. K. (2015). Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1931–1939.

Knopman, D. S., Boeve, B. F., and Petersen, R. C. (2003). Essentials of the proper diagnoses of mild cognitive impairment, dementia, and major subtypes of dementia. In *Mayo Clinic Proceedings*, volume 78, pages 1290–1308. Elsevier.

Kocagoncu, E., Nesbitt, D., Emery, T., Hughes, L. E., Henson, R. N., Rowe, J. B., et al. (2022). Neurophysiological and brain structural markers of cognitive frailty differ from alzheimer's disease. *Journal of Neuroscience*, 42(7):1362–1373.

Koestinger, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2011). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2144–2151. IEEE.

Kok, A. (2022). Cognitive control, motivation and fatigue: A cognitive neuroscience perspective. *Brain and Cognition*, 160:105880.

Królak, A. and Strumiłło, P. (2012). Eye-blink detection system for human–computer interaction. *Universal Access in the Information Society*, 11(4):409–419.

Kuhnke, F. and Ostermann, J. (2019). Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10164–10173.

Kunz, M., Scharmann, S., Hemmeter, U., Schepelmann, K., and Lautenbacher, S. (2007). The facial expression of pain in patients with dementia. *PAIN*®, 133(1-3):221–228.

Kuwabara, S. A., Van Voorhees, B. W., Gollan, J. K., and Alexander, G. C. (2007). A qualitative exploration of depression in emerging adulthood: disorder, development, and social context. *General hospital psychiatry*, 29(4):317–324.

La Cascia, M., Sclaroff, S., and Athitsos, V. (2000). Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE Transactions on pattern analysis and machine intelligence*, 22(4):322–336.

Ladas, A., Frantzidis, C., Bamidis, P., and Vivas, A. B. (2014). Eye blink rate as a biological marker of mild cognitive impairment. *International Journal of Psychophysiology*, 93(1):12–16.

Lagun, D., Manzanares, C., Zola, S. M., Buffalo, E. A., and Agichtein, E. (2011). Detecting cognitive impairment by eye movement analysis using automatic classification algorithms. *Journal of neuroscience methods*, 201(1):196–203.

Laksana, E., Baltrušaitis, T., Morency, L.-P., and Pestian, J. P. (2017). Investigating facial behavior indicators of suicidal ideation. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 770–777. IEEE.

Lalonde, M., Byrns, D., Gagnon, L., Teasdale, N., and Laurendeau, D. (2007). Real-time eye blink detection with gpu-based sift tracking. In *Fourth Canadian Conference on Computer and Robot Vision (CRV'07)*, pages 481–487. IEEE.

Larner, A. (2005). "who came with you?" a diagnostic observation in patients with memory problems? *Journal of Neurology, Neurosurgery & Psychiatry*, 76(12):1739–1739.

Larner, A. (2012). Head turning sign: pragmatic utility in clinical diagnosis of cognitive impairment. *J Neurol Neurosurg Psychiatry*, 83(8):852–853.

Larner, A. J. (2014a). *Dementia in clinical practice: a neurological perspective: pragmatic studies in the cognitive function clinic*. Springer.

Larner, A. J. (2014b). Screening utility of the “attended alone” sign for subjective memory impairment. *Alzheimer Disease & Associated Disorders*, 28(4):364–365.

Larner, A. J. (2018). Number Needed to Diagnose, Predict, or Misdiagnose: Useful Metrics for Non-Canonical Signs of Cognitive Status? *Dementia and Geriatric Cognitive Disorders Extra*, pages 321–327.

Le, V., Brandt, J., Lin, Z., Bourdev, L., and Huang, T. S. (2012). Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer.

Lee, K. H., Algase, D. L., and McConnell, E. S. (2013). Daytime observed emotional expressions of people with dementia. *Nursing research*, 62(4):218.

Lee, K. H., Lee, J. Y., Kim, B., and Boltz, M. (2022). Event-specific emotional expression of persons living with dementia in long-term care: A 6 months follow-up study. *Clinical Nursing Research*, 31(2):320–328.

Lee, M., Lee, Y. K., Lim, M.-T., and Kang, T.-K. (2020). Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features. *Applied Sciences*, 10(10):3501.

Leo, M., Carcagnì, P., Mazzeo, P. L., Spagnolo, P., Cazzato, D., and Distante, C. (2020). Analysis of facial information for healthcare applications: a survey on computer vision-based approaches. *Information*, 11(3):128.

Li, H., He, P., Wang, S., Rocha, A., Jiang, X., and Kot, A. C. (2018a). Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(10):2639–2652.

Li, L. and Feng, X. (2019). Face anti-spoofing via deep local binary pattern. *Deep Learning in Object Detection and Recognition*, pages 91–111.

Li, P., Pei, Y., Zhong, Y., Guo, Y., and Zha, H. (2021). Robust 3d face reconstruction from single noisy depth image through semantic consistency. *IET Computer Vision*, 15(6):393–404.

Li, P., Wang, D., Wang, L., and Lu, H. (2018b). Deep visual tracking: Review and experimental comparison. *Pattern Recognition*, 76:323–338.

Li, Y., Chang, M.-C., and Lyu, S. (2018c). In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. IEEE.

Li, Y., Gong, S., and Liddell, H. (2000). Support vector regression and classification based multi-view face detection and recognition. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 300–305. IEEE.

Li, Y., Gong, S., Sherrah, J., and Liddell, H. (2004). Support vector machine based multi-view face detection and recognition. *Image and Vision computing*, 22(5):413–427.

Little, G., Krishna, S., Black, J., and Panchanathan, S. (2005). A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose angle and illumination angle. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, pages ii–89. IEEE.

Liu, Y., Chen, J., Su, Z., Luo, Z., Luo, N., Liu, L., and Zhang, K. (2016). Robust head pose estimation using dirichlet-tree distribution enhanced random forests. *Neurocomputing*, 173:42–53.

Lövheim, H., Sandman, P.-O., Karlsson, S., and Gustafson, Y. (2009). Sex differences in the prevalence of behavioral and psychological symptoms of dementia. *International psychogeriatrics*, 21(3):469–475.

Luo, B., Shen, J., Wang, Y., and Pantic, M. (2019). The ibug eye segmentation dataset. In *2018 Imperial College Computing Student Workshop (ICCSW 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Ma, B., Huang, R., and Qin, L. (2015). Vod: a novel image representation for head yaw estimation. *Neurocomputing*, 148:455–466.

Mackert, A., Flechtner, K.-M., Woyth, C., and Frick, K. (1991). Increased blink rates in schizophrenics: influences of neuroleptics and psychopathology. *Schizophrenia research*, 4(1):41–47.

Maddage, N. C., Senaratne, R., Low, L.-S. A., Lech, M., and Allen, N. (2009). Video-based detection of the clinical depression in adolescents. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3723–3726. IEEE.

Madrigal, F. and Lerasle, F. (2020). Robust head pose estimation based on key frames for human-machine interaction. *EURASIP Journal on Image and Video Processing*, 2020:1–19.

Maior, C. B. S., das Chagas Moura, M. J., Santana, J. M. M., and Lins, I. D. (2020). Real-time classification for autonomous drowsiness detection using eye aspect ratio. *Expert Systems with Applications*, 158:113505.

Malik, K. and Smolka, B. (2014). Eye blink detection using local binary patterns. In *2014 International Conference on Multimedia Computing and Systems (ICMCS)*, pages 385–390. IEEE.

Martorana, A., Esposito, Z., and Koch, G. (2010). Beyond the cholinergic hypothesis: do current drugs work in alzheimer's disease? *CNS neuroscience & therapeutics*, 16(4):235–245.

Maynard, S. K. (1987). Interactional functions of a nonverbal sign head movement in japanese dyadic casual conversation. *Journal of pragmatics*, 11(5):589–606.

McCann, T. V. and Lubman, D. I. (2012). Young people with depression and their experience accessing an enhanced primary care service for youth with emerging mental health problems: a qualitative study. *BMC psychiatry*, 12(1):1–9.

McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., and Graesser, A. (2007). Facial features for affective state detection in learning environments. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2011). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., Mohs, R. C., Morris, J. C., Rossor, M. N., Scheltens, P., Carrillo, M. C., Thies, B., Weintraub, S., and Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia*, 7(3):263–269.

Mehrabian, A. (2017). Communication without words. In *Communication theory*, pages 193–200. Routledge.

Meyer, G. P., Gupta, S., Frosio, I., Reddy, D., and Kautz, J. (2015). Robust model-based 3d head pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3649–3657.

Mirheidari, B., Blackburn, D., and Christensen, H. (2022). Automatic cognitive assessment: Combining sparse datasets with disparate cognitive scores. *Proc. Interspeech 2022*, pages 2463–2467.

Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M., and Christensen, H. (2017). Toward the automation of diagnostic conversation analysis in patients with memory complaints. *Journal of Alzheimer's Disease*, 58(2):373–387.

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., and Christensen, H. (2018). Dementia detection using automatic analysis of conversations. *Computer Speech & Language*.

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., and Christensen, H. (2019). Dementia detection using automatic analysis of conversations. *Computer Speech & Language*, 53:65–79.

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

Morency, L.-P., Stratou, G., DeVault, D., Hartholt, A., Lhommet, M., Lucas, G., Morbini, F., Georgila, K., Scherer, S., Gratch, J., et al. (2015). Simsensei demonstration: a perceptive virtual human interviewer for healthcare applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Mota, I. A. and Lins, O. G. (2017). Bereitschaftspotential preceding spontaneous and voluntary eyelid blinks in normal individuals. *Clinical neurophysiology*, 128(1):100–105.

Muliyala, K. P. and Varghese, M. (2010). The complex relationship between depression and dementia. *Annals of Indian Academy of Neurology*, 13(Suppl2):S69.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Murphy-Chutorian, E., Doshi, A., and Trivedi, M. M. (2007). Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *2007 IEEE intelligent transportation systems conference*, pages 709–714. IEEE.

Nakano, T. (2015). Blink-related dynamic switching between internal and external orienting networks while viewing videos. *Neuroscience Research*, 96:54–58.

Nanthini, N., Puviarasan, N., and Aruna, P. (2022). Eye blink-based liveness detection using odd kernel matrix in convolutional neural networks. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 1*, pages 473–483. Springer.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., and Chertkow, H. (2005). The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.

Navastara, D. A., Putra, W. Y. M., and Fatichah, C. (2020). Drowsiness detection based on facial landmark and uniform local binary pattern. In *Journal of Physics: Conference Series*, volume 1529, page 052015. IOP Publishing.

Nemeroff, C. B. (2007). The burden of severe depression: a review of diagnostic challenges and treatment alternatives. *Journal of psychiatric research*, 41(3-4):189–206.

O'Malley, R. P. D., Mirheidari, B., Harkness, K., Reuber, M., Venneri, A., Walker, T., Christensen, H., and Blackburn, D. (2021). Fully automated cognitive screening tool based on assessment of speech and language. *Journal of Neurology, Neurosurgery & Psychiatry*, 92(1):12–15.

Ouanan, H., Ouanan, M., and Aksasse, B. (2016). Facial landmark localization: Past, present and future. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pages 487–493.

Oyama, A., Takeda, S., Ito, Y., Nakajima, T., Takami, Y., Takeya, Y., Yamamoto, K., Sugimoto, K., Shimizu, H., Shimamura, M., et al. (2019). Novel method for rapid assessment of cognitive impairment using high-performance eye-tracking technology. *Scientific reports*, 9(1):12932.

Pampouchidou, A., Marias, K., Tsiknakis, M., Simos, P., Yang, F., Lemaître, G., and Meriaudeau, F. (2016). Video-based depression detection using local curvelet binary patterns in pairwise orthogonal planes. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3835–3838. IEEE.

Pampouchidou, A., Simos, P. G., Marias, K., Meriaudeau, F., Yang, F., Pediaditis, M., and Tsiknakis, M. (2017). Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions on Affective Computing*, 10(4):445–470.

Pan, G., Sun, L., Wu, Z., and Lao, S. (2007). Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE.

Papakostas, M., Spyrou, E., Giannakopoulos, T., Siantikos, G., Sgouropoulos, D., Mylonas, P., and Makedon, F. (2017). Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. *Computation*, 5(2):26.

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

Pereira, M. L. F., Marina von Zuben, A. C., Aprahamian, I., and Forlenza, O. V. (2014). Eye movement analysis and cognitive processing: detecting indicators of conversion to alzheimer's disease. *Neuropsychiatric disease and treatment*, 10:1273.

Pérez Espinosa, H., Escalante, H. J., Villaseñor-Pineda, L., Montes-y Gómez, M., Pinto-Avedaño, D., and Reyez-Meza, V. (2014). 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*, pages 49–55.

Petersen, R. C., Lopez, O., Armstrong, M. J., Getchius, T. S., Ganguli, M., Gloss, D., Gronseth, G. S., Marson, D., Pringsheim, T., Day, G. S., et al. (2018). Practice guideline update summary: Mild cognitive impairment: Report of the guideline development, dissemination, and implementation subcommittee of the american academy of neurology. *Neurology*, 90(3):126–135.

Petti, U., Baker, S., and Korhonen, A. (2020). A systematic literature review of automatic alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, 27(11):1784–1797.

Ponder, E. and Kennedy, W. (1927). On the act of blinking. *Quarterly journal of experimental physiology: Translation and integration*, 18(2):89–110.

Portello, J. K., Rosenfield, M., and Chu, C. A. (2013). Blink rate, incomplete blinks and computer vision syndrome. *Optometry and Vision Science*, 90(5):482–487.

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., and O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2:927312.

Pusdekar, S. J. and Chhaware, S. P. (2014). Using visual clues concept for extracting main data from deep web pages. In *2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies*, pages 190–193. IEEE.

Qin, J., Shimoyama, T., and Sugano, Y. (2022). Learning-by-novel-view-synthesis for full-face appearance-based 3d gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4991.

Radlak, K., Bozek, M., and Smolka, B. (2015). Silesian deception database: Presentation and analysis. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 29–35.

Radlak, K. and Smolka, B. (2012). A novel approach to the eye movement analysis using a high speed camera. In *2012 2nd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*, pages 145–150. IEEE.

Radlak, K. and Smolka, B. (2013). Blink detection based on the weighted gradient descriptor. In *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013*, pages 691–700. Springer.

Rahmaniar, W., ul Haq, Q. M., and Lin, T.-L. (2022). Wide range head pose estimation using a single rgb camera for intelligent surveillance. *IEEE Sensors Journal*, 22(11):11112–11121.

Ranjan, R., Patel, V. M., and Chellappa, R. (2017). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):121–135.

Redondo-Cabrera, C., López-Sastre, R. J., Xiang, Y., Tuytelaars, T., and Savarese, S. (2016). Pose estimation errors, the ultimate diagnosis. In *European Conference on Computer Vision*, pages 118–134. Springer.

Richmond, V. P., McCroskey, J. C., and Hickson, M. (2008). *Nonverbal behavior in interpersonal relations*. Allyn & Bacon.

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

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.-M., et al. (2019). Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 3–12. ACM.

Rodriguez, J. D., Ousler III, G. W., Johnston, P. R., Lane, K., and Abelson, M. B. (2013). Investigation of extended blinks and interblink intervals in subjects with and without dry eye. *Clinical Ophthalmology (Auckland, NZ)*, 7:337.

Rosten, E. and Drummond, T. (2006). Machine learning for high-speed corner detection. In *European conference on computer vision*, pages 430–443. Springer.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283.

Sagar, R. (2021). Big data to good data: Andrew ng urges ml community to be more data-centric and less model-centric. *Analytics India Magazine. Accessed May*, 17:2021.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18.

Sagonas, C., Panagakis, Y., Zafeiriou, S., and Pantic, M. (2014). Raps: Robust and efficient automatic construction of person-specific deformable models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1789–1796.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2013a). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 397–403.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2013b). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 397–403.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2013c). A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 896–903.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229.

Sanderson, C. and Paliwal, K. K. (2004). Identity verification using speech and face information. *Digital Signal Processing*, 14(5):449–480.

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

Schillingmann, L. and Nagai, Y. (2015). Yet another gaze detector: An embodied calibration free system for the icub robot. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 8–13. IEEE.

Schmidtke, K., Pohlmann, S., and Metternich, B. (2008). The syndrome of functional memory disorder: definition, etiology, and natural course. *The American Journal of Geriatric Psychiatry*, 16(12):981–988.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Seblova, D., Brayne, C., Mach, V., Kuklová, M., Kopecek, M., and Cermakova, P. (2019). Changes in cognitive impairment in the czech republic. *Journal of Alzheimer's Disease*, 72(3):693–701.

Seha, S., Papangelakis, G., Hatzinakos, D., Zandi, A. S., and Comeau, F. J. (2019). Improving eye movement biometrics using remote registration of eye blinking patterns. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2562–2566. IEEE.

Senoussaoui, M., Sarria-Paja, M., Santos, J. F., and Falk, T. H. (2014). Model fusion for multimodal depression classification and level detection. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 57–63.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., and Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 50–58.

Siebert, C. F. and Siebert, D. C. (2017). *Data analysis with small samples and non-normal data: Nonparametrics and other strategies*. Oxford University Press.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.

Smith, M. (1995). Facial expression in mild dementia of the alzheimer type. *Behavioural Neurology*, 8:149–156.

Smith, M., Robinson, L., and Segal, R. (2022). Age-related memory loss. Accessed on February 6, 2023.

Song, F., Tan, X., Liu, X., and Chen, S. (2014). Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients. *Pattern Recognition*, 47(9):2825–2838.

Soukupová, T. and Cech, J. (2016). Real-time eye blink detection using facial landmarks. In *21st Computer Vision Winter Workshop*.

Soysal, P., Usarel, C., Ispirli, G., and Isik, A. T. (2017). Attended with and head-turning sign can be clinical markers of cognitive impairment in older adults. *International psychogeriatrics*, 29(11):1763–1769.

Stern, J. A., Walrath, L. C., and Goldstein, R. (1984). The endogenous eyeblink. *Psychophysiology*, 21(1):22–33.

Stratou, G., Scherer, S., Gratch, J., and Morency, L.-P. (2013). Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 147–152. IEEE.

Sun, W. S., Baker, R. S., Chuke, J. C., Rouholiman, B. R., Hasan, S. A., Gaza, W., Stava, M. W., and Porter, J. D. (1997). Age-related changes in human blinks. passive and active changes in eyelid kinematics. *Investigative ophthalmology & visual science*, 38(1):92–99.

Taati, B., Zhao, S., Ashraf, A. B., Asgarian, A., Browne, M. E., Prkachin, K. M., Mihailidis, A., and Hadjistavropoulos, T. (2019). Algorithmic bias in clinical populations—evaluating and improving facial analysis technology in older adults with dementia. *IEEE access*, 7:25527–25534.

Tanaka, H., Adachi, H., Kazui, H., Ikeda, M., Kudo, T., and Nakamura, S. (2019). Detecting dementia from face in human-agent interaction. In *Adjunct of the 2019 International Conference on Multimodal Interaction*, pages 1–6.

Tanaka, H., Adachi, H., Ukita, N., Ikeda, M., Kazui, H., Kudo, T., and Nakamura, S. (2017). Detecting dementia through interactive computer avatars. *IEEE journal of translational engineering in health and medicine*, 5:1–11.

Tanaka, H., Adachi, H., Ukita, N., Kudo, T., and Nakamura, S. (2016). Automatic detection of very early stage of dementia through multimodal interaction with computer avatars. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 261–265.

Taylor, J., Elsworth, J., Lawrence, M., Sladek Jr, J., Roth, R., and Redmond Jr, D. (1999). Spontaneous blink rates correlate with dopamine levels in the caudate nucleus of mptp-treated monkeys. *Experimental neurology*, 158(1):214–220.

Tomasi, C. and Kanade, T. (1991). Detection and tracking of point. *Int J Comput Vis*, 9:137–154.

Travis, L. B., Roberts, G. D., and Wilson, W. R. (1985). Clinical significance of pseudallescheria boydii: a review of 10 years' experience. In *Mayo Clinic Proceedings*, volume 60, pages 531–537. Elsevier.

Tukey, J. W. et al. (1977). *Exploratory data analysis*, volume 2. Reading, MA.

Tyson, B., Cabrera, L., Scriven, E., Larios, C., Reilly, E., and Kearns, L. (2019). The diagnostic utility of the "attended alone" sign for dementia in patients presenting for neuropsychological evaluation. *J Neuroinflamm Neurodegener Dis*, 3(1):100010.

Tzimiropoulos, G., Alabort-i Medina, J., Zafeiriou, S., and Pantic, M. (2012). Generic active appearance models revisited. In *Asian Conference on Computer Vision*, pages 650–663. Springer.

UCSF Memory and Aging Center (2020). Healthy ageing.

Utaminingrum, F., Purwanto, A. D., Masruri, M. R. R., Ogata, K., and Somawirata, I. K. (2021). Eye movement and blink detection for selecting menu on-screen display using probability analysis based on facial landmark. *International Journal of Innovative Computing, Information and Control*, 17(4):1287–1303.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., and Pantic, M. (2014). Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. (2013). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM.

van de Mortel, L. A., Thomas, R. M., van Wingen, G. A., Initiative, A. D. N., et al. (2021). Grey matter loss at different stages of cognitive decline: A role for the thalamus in developing alzheimer's disease. *Journal of Alzheimer's Disease*, 83(2):705–720.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

Volkow, N. D., Gur, R. C., Wang, G.-J., Fowler, J. S., Moberg, P. J., Ding, Y.-S., Hitzemann, R., Smith, G., and Logan, J. (1998). Association between decline in brain dopamine activity with age and cognitive and motor impairment in healthy individuals. *American Journal of psychiatry*, 155(3):344–349.

von Cramon, D. and Schuri, U. (1980). Blink frequency and speech motor activity. *Neuropsychologia*, 18(4-5):603–606.

Wakefield, S. J., Blackburn, D. J., Harkness, K., Khan, A., Reuber, M., and Venneri, A. (2018). Distinctive neuropsychological profiles differentiate patients with functional memory disorder from patients with amnestic-mild cognitive impairment. *Acta Neuropsychiatrica*, 30(2):90–96.

Walker, G., Walker, T., O'Malley, R., Mirheidari, B., Christensen, H., Reuber, M., and Blackburn, D. (2023). Features of answers to questions about recent events by people with mild cognitive impairment and alzheimer's disease, and healthy controls. *Journal of Interactional Research in Communication Disorders*.

Wang, M., Guo, L., and Chen, W.-Y. (2017). Blink detection using adaboost and contour circle for fatigue recognition. *Computers & Electrical Engineering*, 58:502–512.

Wei, Y., Varanasi, R. S., Schwarz, T., Gomell, L., Zhao, H., Larson, D. J., Sun, B., Liu, G., Chen, H., Raabe, D., et al. (2021). Machine-learning-enhanced time-of-flight mass spectrometry analysis. *Patterns*, 2(2):100192.

Werner, P., Al-Hamadi, A., Limbrecht-Ecklundt, K., Walter, S., and Traue, H. C. (2018). Head movements and postures as pain behavior. *PloS one*, 13(2):e0192767.

WHO (2022). Dementia.

Wilcockson, T. D., Mardanbegi, D., Xia, B., Taylor, S., Sawyer, P., Gellersen, H. W., Leroi, I., Killick, R., and Crawford, T. J. (2019). Abnormalities of saccadic eye movements in dementia due to alzheimer's disease and mild cognitive impairment. *Aging (Albany NY)*, 11(15):5389.

Woodard, J. L., Seidenberg, M., Nielson, K. A., Antuono, P., Guidotti, L., Durgerian, S., Zhang, Q., Lancaster, M., Hantke, N., Butts, A., et al. (2009). Semantic memory activation in amnestic mild cognitive impairment. *Brain*, 132(8):2068–2078.

Woodruff-Pak, D. S. (2001). Eyeblink classical conditioning differentiates normal aging from alzheimer's disease. *Integrative Physiological & Behavioral Science*, 36(2):87–108.

Wu, Y., Gou, C., and Ji, Q. (2017). Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3471–3480.

Xia, H., Liu, G., Xu, L., and Gan, Y. (2022). Collaborative learning network for head pose estimation. *Image and Vision Computing*, 127:104555.

Xiao, S., Yan, S., and Kassim, A. A. (2015). Facial landmark detection via progressive initialization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 33–40.

Xiong, X. and De la Torre, F. (2015). Global supervised descent method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2664–2673.

Yang, J., Deng, J., Zhang, K., and Liu, Q. (2015). Facial shape tracking via spatio-temporal cascade shape regression. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 41–49.

Yang, J., Rahardja, S., and Fränti, P. (2018). Mean-shift outlier detection. In *FSDM*, pages 208–215.

Yang, J., Rahardja, S., and Fränti, P. (2019). Outlier detection: how to threshold outlier scores? In *Proceedings of the international conference on artificial intelligence, information processing and cloud computing*, pages 1–6.

Yu, Y., Mora, K. A. F., and Odobez, J.-M. (2017). Robust and accurate 3d head pose estimation through 3dmm and online head model reconstruction. In *2017 12th ieee international conference on automatic face & gesture recognition (fg 2017)*, pages 711–718. Ieee.

Zadeh, A., Chong Lim, Y., Baltrusaitis, T., and Morency, L.-P. (2017). Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2519–2528.

Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., and Shen, J. (2017). The menpo facial landmark localisation challenge: A step towards the solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 170–179.

Zhao, L., Wang, Z., Zhang, G., Qi, Y., and Wang, X. (2018). Eye state recognition based on deep integrated neural network and transfer learning. *Multimedia Tools and Applications*, 77(15):19415–19438.

Zheng, J., Chu, H., Struppa, D., Zhang, J., Yacoub, S. M., El-Askary, H., Chang, A., Ehwerhemuepha, L., Abudayyeh, I., Barrett, A., et al. (2020). Optimal multi-stage arrhythmia classification approach. *Scientific reports*, 10(1):1–17.

Zhou, D., Luo, J., Silenzio, V. M., Zhou, Y., Hu, J., Currier, G., and Kautz, H. (2015). Tackling mental health by integrating unobtrusive multimodal sensing. In *Twenty-ninth AAAI conference on artificial intelligence*.

Zhou, X., Huang, P., Liu, H., and Niu, S. (2019). Learning content-adaptive feature pooling for facial depression recognition in videos. *Electronics Letters*, 55(11):648–650.

Zhou, X., Jin, K., Shang, Y., and Guo, G. (2018). Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing*, 11(3):542–552.

Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE.

Zhu, Y., Shang, Y., Shao, Z., and Guo, G. (2017). Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Transactions on Affective Computing*, 9(4):578–584.

Zolfaghari, S., Khodabandehloo, E., and Riboni, D. (2022). Traminer: Vision-based analysis of locomotion traces for cognitive assessment in smart-homes. *Cognitive Computation*, 14(5):1549–1570.

Appendix A

Reviewing Previous Work's Datasets

The following section reviews two types of datasets: commonly used datasets to evaluate state-of-art techniques (e.g., face detection and tracking, eye blink detection and head movement estimation) and health conditions datasets. Each dataset is briefly described, including its purpose, whether it is in the wild according to the authors or not, the number of the population if mentioned, the number of images or videos, challenges, limitations and availability.

A.1 Commonly Used Data

A.1.1 Face Detection and Tracking

- **The LFW dataset** is collected for investigating face recognition in unconstrained environments. The authors refer to it as an in-the-wild dataset with 13,233 images of 5,749 famous figures. The images include a few children, very few people older than 80, and a few women. Several conditions, such as poor illumination, large head pose, low resolution and other vital factors, represent a minor part of LFW. This dataset is limited in the representation of some groups; there are no babies and only a small proportion of some ethnicities. In addition, it does not contain images with strong occlusion. More importantly, the authors mentioned that the size of this dataset is small to consider as evidence of good performance in evaluating a particular approach. The dataset is available at http://vis-www.cs.umass.edu/lfw/.

- **The Helen dataset** is collected for facial component localisation. It is in-the-wild data, according to the authors, with 2330 images. It is constructed in several steps: selecting 2330 images with high resolution and large variations and including many people from different cultures. Then, a face detector is used to filter the selected subset to include images with large faces (i.e., greater than 500 pixels in width). The subset is manually filtered further to remove profile views, false positives and low-quality images. For each image, a cropped version is created to include the face and a small proportion of the background. The face is not always at the centre of the image. A cropped image may contain more than one person. Every image is annotated manually to localise the eyes, nose, mouth, eyebrows and jaw-line using Amazon Mechanical Turk. This dataset has several limitations, such as using multiple filtering approaches to exclude false positives, profile views, faces less than 500 pixels in width, and low-resolution images. It is available at http://www.ifp.illinois.edu/~vuongle2/helen/.

- **The AFW dataset** is built for face detection, pose estimation and facial landmarks detection. According to the authors, this data is considered in the wild and consists of 250 images with 468 faces. It is collected from Flicker images with variations in the background, face poses and appearances, such as ageing, makeup, skin colour, facial expressions and sunglasses. It is annotated in terms of the faces' bounding box, poses and landmarks. For each face, a bounding box is determined, six facial landmark points are annotated (i.e., the centre of the eyes, the tip of the nose, the two corners and the centre of the mouth), and the yaw and pitch angles of each pose are added. The data is limited to the size of the dataset and, more importantly, the lack of information in terms of the challenges in this dataset. The dataset is available at https://datasets.activeloop.ai/docs/ml/datasets/afw-dataset/.

- **The LFPW dataset** is collected for facial landmarks localisation. According to the authors, it is in-the-wild data with 3000 images of faces collected from the Internet, particularly Google, Flickr and Yahoo. These 3000 faces were detected. Then, any image incorrectly detected or near-profile face was excluded from the subset. Finally, every image was annotated with 35 points on each face by Amazon Mechanical Turk. The images include faces where hair, sunglasses or glasses may occlude the eye, and a hat, a cigarette, a hand, or a microphone may cover some parts of the face. Some facial landmarks may be occluded by facial hair, and some may exhibit a strong shadow across some face parts. There are images with facial expressions, and faces can have makeup, be made up theatrically or no makeup. This dataset's limitations are similar to those mentioned for the Helen dataset. For example, a filtering approach to exclude false positives, profile views or near profile views, and low-quality images. A sample of this dataset is available at https://neerajkumar.org/databases/lfpw/.

- **The 300-W face dataset** is collected to measure the facial landmark detectors' feasibility in handling in-the-wild conditions. According to the authors, it is in-the-wild data, comprising 300 indoor and 300 outdoor images. It is gathered from the Internet using tags such as ’party’, ’conference’, ’protests’, ’football’, and ’celebrities’. This data includes a higher proportion of partially-occluded faces than the other datasets. It covers more facial expressions, such as ’neutral’, ’smile’, ’surprise’, or ’scream’, compared with other in-the-wild datasets, such as Helen and LFPW, that include only ’smile’ expressions. In addition, it consists of a similar pose proportion to that in AFW and variations in background, illumination and image quality. The ground truth is created using a semi-automatic method to annotate facial landmarks, following previous work (Sagonas et al., 2013c; Tzimiropoulos et al., 2012). This dataset is limited in the number of partially occluded faces and the small number of images used for testing. It is available at https://ibug.doc.ic.ac.uk/resources/300-W/.

- **The IJB-FL dataset** is a subset of IJB-A (Klare et al., 2015), which is a benchmark for face recognition and consists of variation in image conditions, ethnicities and full poses. This data is referred to as in-the-wild data. The authors took a sample of 180 images from IJB-A, of which 128 are frontal and 52 are profile. This dataset is annotated manually with about 68 facial landmarks depending on the face visibility. This data differs from the previous ones because it includes several images in a non-frontal pose. It is also limited by the lack of information regarding its challenges, which would be helpful for comparing it with other datasets. The dataset is available at http://face.nist.gov.

- **The Menpo dataset** is constructed for facial landmarks localisation and addresses the limitations of previous in-the-wild datasets (i.e., 300W and 300-VW, which will be described later). The limitations of the previous datasets are that they include few images with faces in extreme poses, and the test set consists of a very small number of images, about 600. Hence, Menpo data includes a training set with 5658 semi-frontal and 1906 profile images and a test set with 5335 frontal and 1946 profile images. The profile images are annotated manually with up to 39 landmarks. For semi-frontal images, the annotation has been conducted using a semi-automatic approach. Finally, the faces in the images are cropped using the landmarks and provided for training and test purposes. This data is also limited by the lack of information in terms of its challenges, which would be helpful for comparing it with other datasets. It is available at https://ibug.doc.ic.ac.uk/resources/2nd-facial-landmark-tracking-competition-menpo-ben/.

- **The RU-FACS dataset** is collected for the facial action unit recognition task. It contains 33 videos of different participants recorded in a lab-controlled environment. Each video lasts for about 2 minutes. This dataset is limited by the lack of information in terms of the challenges in this dataset, which could be helpful to compare with other datasets.

- **The YouTube Celebrities dataset** is collected for face tracking and recognition task. The authors claimed that the data is from real-world scenarios, so it can be referred to as an in-the-wild dataset, consisting of 1910 sequences of 47 celebrities in an interview or TV show. It includes variations in face pose, occlusion and illumination. It is captured at 25 fps, which is the reason for including videos with low resolution. Most of the videos are less than 15 seconds long. This dataset has the same limitation as the previous datasets, which is the lack of information in terms of the challenges in it, which would be helpful for comparing it with other datasets. It is available at http://seqamlab.com/youtube-celebrities-face-tracking-and-recognition-dataset/.

- **The 300-VW dataset** is built for the facial landmark tracking task. The authors claimed that it is in-the-wild, and it includes 300 videos collected mostly from YouTube and 12 clips from the SEMAINE database (McKeown et al., 2011). Each video shows only one person and is captured at 30 fps, and the number of frames in total is 218559. The average duration of the videos is 64 seconds. The authors divided the data into three categories. Category one contains videos of people in good-environmental conditions, various head poses, and occlusions by glasses or beards. Category two includes videos of people recorded in various illuminations and dark rooms, and displaying facial expressions with small head poses. Category three consists of videos recorded in unconstrained conditions, such as variations in illumination, occlusions, facial expressions, head poses and makeup. This dataset is limited by the lack of information regarding the participants' behaviour, look, demographics, and background. It is available at https://ibug.doc.ic.ac.uk/resources/300-VW/.

- **The DDF dataset** is constructed for face tracking from profile to profile. The authors did not state whether the data is in-the-wild or not. The dataset contains 15 videos with a total of 10,822 frames. Each video displays one participant pretending to be distracted while driving in a stationary vehicle or indoor environment. 12 videos of 15 are recorded with participants sitting inside a vehicle; five of them participants were recorded at nighttime and under infrared (IR) light, and the other seven were recorded under natural lighting. The remaining three participants are recorded indoors. This dataset is also limited by the lack of information about its challenges.

- **The NDS dataset** is also constructed for face tracking from profile to profile. Again, the authors did not state whether this data is in-the-wild or not. It consists of 20 subsequences of drivers' faces recorded during a drive between two areas, the Blacksburg, VA and Washington, DC, areas. It is considered more challenging than the DDF dataset due to the videos' lower spatial and temporal resolution. Each video lasts for one minute and is captured at 15 fps with a resolution of 360 x 240. Both the DDF and NDS datasets include many faces in near-frontal or profile faces (±90° yaw, ±50° pitch), many frames under extreme IR lighting conditions and occlusion by sunglasses. However, this data is still limited by the lack of information about its challenges.

A.1.2 Eye Blink Detection

- **The ZJU dataset** is recorded in a lab-controlled environment and includes 80 videos of 20 participants, each lasting only a few seconds. For each participant, there are four clips: frontal view without glasses, frontal with thin and black frame glasses, and upward view without glasses. The participants were asked to blink spontaneously at normal speed. The videos are collected by LogitechPro5000, a generic web camera, and captured at 30 fps. There are very short blinks, which last for two frames, indicating that the videos are not captured using 30 fps the whole time. The participants are still and without any head movement.

- **The Talking face dataset** is recorded in a lab-controlled environment and consists of one video of one participant sitting and talking in front of the camera. The participant shows one facial expression, smiling, but without any head or body movements. The background was a blue wall. The video is captured at 25 fps for a total of 2:46 minutes.

- **In the Eyeblink8 dataset**, although the videos are recorded in a home environment, the participants are still and only show smiling expressions. It consists of eight videos of four participants (one wearing glasses). The videos are collected by a Logitech C905 camera at 30fps.

- **The Basler5 dataset** is recorded in a lab-controlled environment. It consists of 5 participants who were sitting at a close distance to the camera and in a lab-controlled environment. The videos are collected by a high-speed Basler camera at 100 fps.

- **The Researcher's Night dataset** is recorded at an event called Researcher's Night 2014, at which people are asked to read an article or blink during the recording. It is collected in a cluttered environment and has 107 videos of different people. The videos are captured at 15 and 30 fps. The recordings include around 20% of people wearing glasses, a little head movement or touching their faces.

A.1.3 Head Movement Estimation

- **The BU dataset** is recorded in a lab-controlled environment and includes only 200 images for five participants, which is the major disadvantage of this dataset. The collection is conducted in two sessions: one in which the illumination is uniform and another in which the illumination varied with complex scenarios regarding the appearance of the face. Head rotations are recorded using a magnetic tracker attached to each participant's head as aground truth, whilst the test set was recorded via a Sony Handy-cam. Both the quality of the images and the annotations are considered to be low. This dataset is limited to frontal faces and other important factors (e.g., eyeglasses, facial hair, etc.). In addition, with some furniture and computers in a typical lab-controlled environment, some approaches showed a lower accuracy, which indicates that a noisy background could affect the performance of a model.

- **The AFLW dataset** is designed for facial landmarks localisation and head movement tracking. According to the authors, it is considered a challenging dataset collected from the Internet in nine different lighting conditions and in the wild. It consists of 21,997 images for 25,993 faces. This dataset includes frontal and non-frontal images showing different facial expressions, variations in face appearance and environment-related factors. The facial landmarks and the head poses are manually annotated. However, no annotation is provided when a facial landmark is not visible. In addition, the annotations provided and the image quality are low. It is available on request at http://lrs.icg.tugraz.at/research/aflw/.

- **The ICT-3DHP dataset** is collected using the Kinect device for the test set and the ground truth is recorded using a magnetic tracker attached to each participant's head. This dataset, therefore, includes both RGB and depth data. It consists of only 10 participants, and the number of videos is 10, which is about 1400 frames. The head rotations for the three angles are recorded. It is then annotated using a Polhemus FASTRAK flock of birds tracker, which results in low-quality in the ground-truth annotation. They mentioned that their dataset is limited to a few images with roll angle. There is no information about the dataset challenges or the recording environment.

- **The BIWI dataset** is collected in a lab-controlled environment based on a recording setting, which involved the participants sitting in front of a Kinect camera at a distance of 1 meter indoors in a living room environment. The recordings contain 20 participants (4 people with glasses). The total number of frames is 15,000. The participants turn their heads to exhibit all possible yaw and pitch angles. The head orientation for these angles varies (e.g., yaw with ±75°, pitch with ±60°, and roll with ±50°). The head movement data are automatically annotated, and this lead to medium-quality of ground-truth annotation is medium. This data is limited to only a few images with roll angles and lacks more information about the participants and the environment.

A.2 Health Conditions Data

- **The BlackDog dataset** is collected by the BlackDog Institute, an organisation focusing on clinical research in Australia. A total of 80 participants are recorded with ages rangeing from 21 to 75, but only 60 of them are used in order to reduce the variability by including only English speakers. The participants are interviewed and asked questions from eight groups of questions, requiring them to explain and describe happy and sad events. Only the participants who met the criteria of Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) (Carter, 2014) diagnostic rules were chosen.

- **In the ORI dataset**, video recordings of eight participants are selected with ages ranging between 12 and 19 years. It includes a one-hour video recording for every participant based on three family interactions: event planning, problem-solving and family consensus interaction. According to the authors, all of the selected participants are white, and none wore eyeglasses during the recording sessions.

- **The AVEC2013 dataset** is an audio-visual depression dataset recorded using a human-computer interaction for people performing several tasks. It includes 340 videos for 292 participants with ages ranging from 18 to 63 years, with an average of 31.5 years. It is recorded using a webcam and a microphone and is captured at 30 fps with a resolution of 640 x 480. Only one person appears in each clip, and some participants recorded more than one clip. The dataset organiser provided only part of this dataset, 150 video clips, split equally into training, development and test sets. This dataset is available for researchers at this location http://avec2013-db.sspnet.eu.

- **The AVEC2014 dataset** is selected partially from the AVEC2013 dataset. It includes 300 videos for 84 participants. Only one person appears in each clip and some participants recorded more than one clip. The dataset consists of only two tasks: Freeform (answering questions) and Northwind (reading a passage). Each task has three partitions: 50 clips for training, 50 for development and 50 for the test.

- **The DAIC-WoZ database** is collected by four interviews based on semi-structured clinical interactions: Face-to-Face, Teleconference, Wizard-of-Oz and Automated. The data consists of 189 audio and video clips and physiological data, such as galvanic skin response, electrocardiogram, and respiration. The verbal and non-verbal features are annotated for this data. The text modality is also collected during the sessions. It is available for researchers like the AVEC2013 and AVEC2014 datasets. The data can be found at http://dcapswoz.ict.usc.edu.

- **The CHI-MEI database** consists of audio and video modalities collected by a clinician for 26 participants in the CHI-MEI Medical Center, Taiwan. It is collected based on six discrete videos designed in order to arouse the participant's ability to express their emotions through their faces and speech answers, covering disgust, fear, sadness, surprise, anger, and happiness.

- **The Pittsburgh dataset** includes 57 participants (34 females and 23 males) with age ranges from 19 to 65 years with a mean of 39.65 years. This data is recorded during clinical treatment for depression, and the participants had to meet the criteria for Major Depressive Disorder, which is assessed for each participant at 1, 7, 13, and 21 weeks. This data is open access for researchers, but details of only 49 of the participants are available for public use due to changes in the original diagnosis and missing clips or audio.

- **The BD dataset** consists of 95 participants (46 patients and 49 healthy controls) with ages ranging from 18 to 60 years. The data is recorded based during semi-structured interviews by the SKIP-TURK. Two different measurements (i.e., the young mania rating scale (YMRS) and the 10-item Montgomery—Asberg depression rating scale (MADRS)) are used to assess the depressive features on particular days (0, 3, 7, 14 and 28). Each day, the audio and videos are recorded, and then each video clip and audio are labelled by YMRS/MADRS ratings. This data is adopted in AVEC2018.

- **The OU2016 dataset** consists of 20 participants. Ten are recorded at the Osaka University Hospital (the dementia group), and the other ten are recorded at the Nara Institute of Science and Technology (healthy controls). Participants with dementia are diagnosed at a very early stage of dementia by expert doctors based on the DSM-IV (Carter, 2014) at the hospital of Osaka University. The data is recorded using a laptop (Surface Pro 3), and the distance between the laptop and the participant was constant. For analysis and research, two participants are removed due to diagnosis issues, resulting in 18 participants (9 with dementia and 9 without).

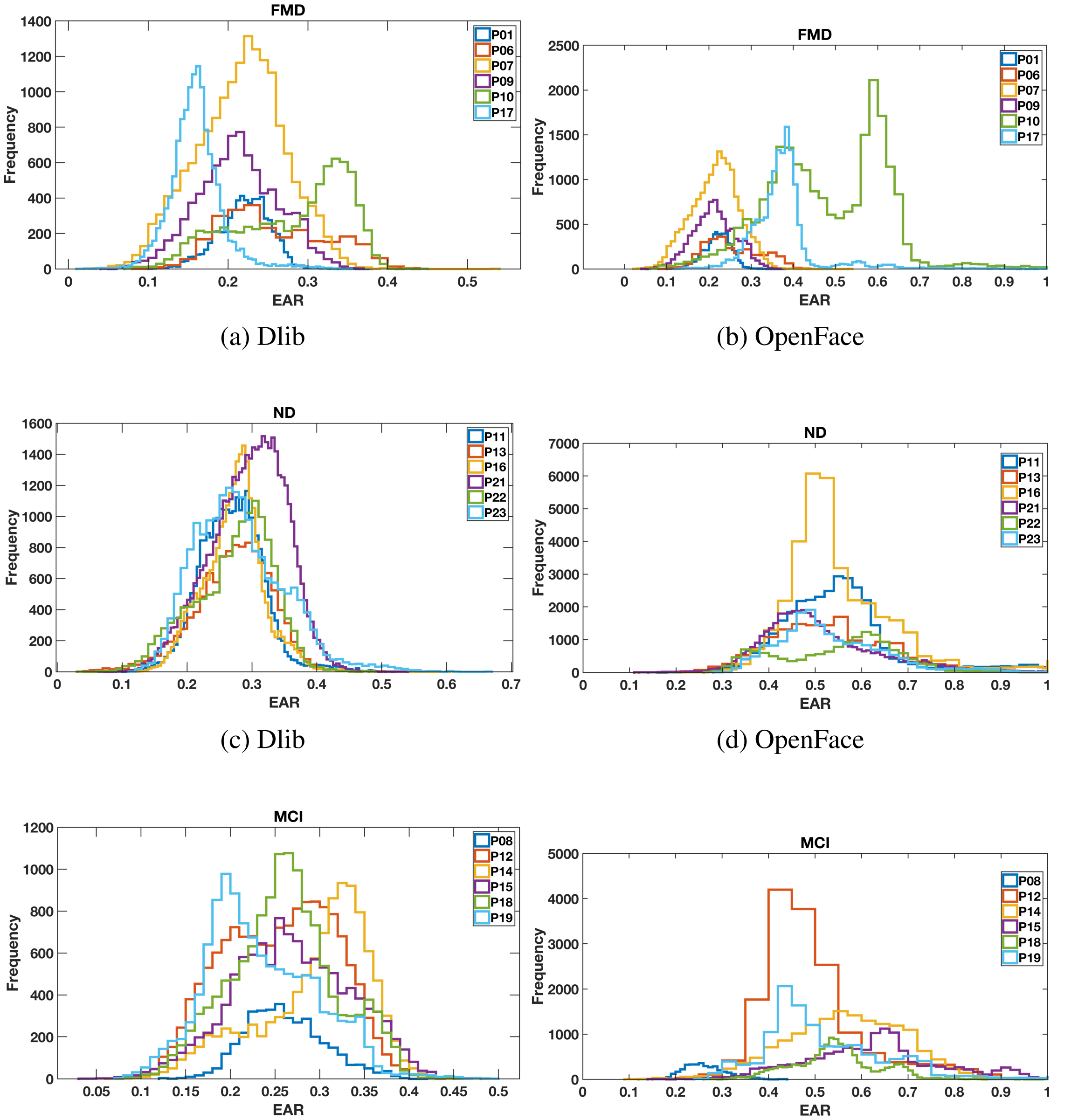
- **In the OU2017 dataset**, 33 Japanese participants (16 with dementia and 17 healthy controls) are recruited. Prior to the recording, the participants sign a consent form. 29 participants are selected, but four were excluded due to diagnosis problems. The dementia group includes participants with mild cognitive impairment (MCI) or dementia who have been diagnosed by an expert clinician at the Osaka University Hospital.
- **The OU2019 dataset** includes 24 participants (12 with dementia and 12 without) with an age average of about 75 years. The dementia group includes Alzheimer's disease (AD), normal pressure hydrocephalus (NPH), one MCI, and AD+NPH. Psychiatrists diagnose the participants at the Osaka University Hospital.

Appendix B

Dlib vs. OpenFace

B.1 The Distribution of the EAR

Figure B.1 presents the distribution of the EAR values, which are calculated using the extracted eyes landmarks of Dlib and OpenFace. Figures B.1a B.1c and B.1e show that the ranges reached 0.55, 0.65 and 0.5, respectively, whereas Figures B.1b B.1d and B.1f shows that the ranges reached 2.5, 33.85 and 13, respectively, on the x-axis. However, the range was cut off to reach a maximum of one to make the histograms clear. It shows that the participants with dementia or MCI have higher values of the EAR on the x-axis, which indicates head movements or turns. Figures B.1c and B.1d show that the mean of every participant on the histogram was close to others, making the histograms overlap. The participants in Figures B.1e and B.1f had only a small difference in the mean between them. By contrast, the histograms for Figures B.1a and B.1b show a large difference in the mean between the participants for Dlib and OpenFace on the histograms.



(a) Dlib

(b) OpenFace

(c) Dlib

(d) OpenFace

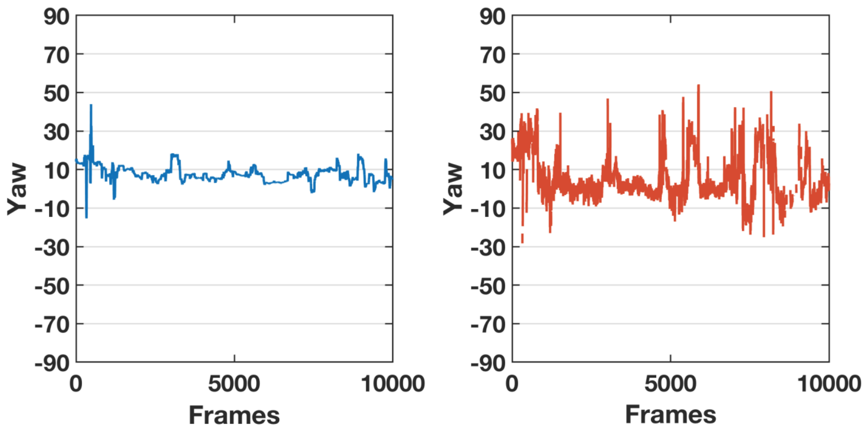
(e) Dlib

(f) OpenFace

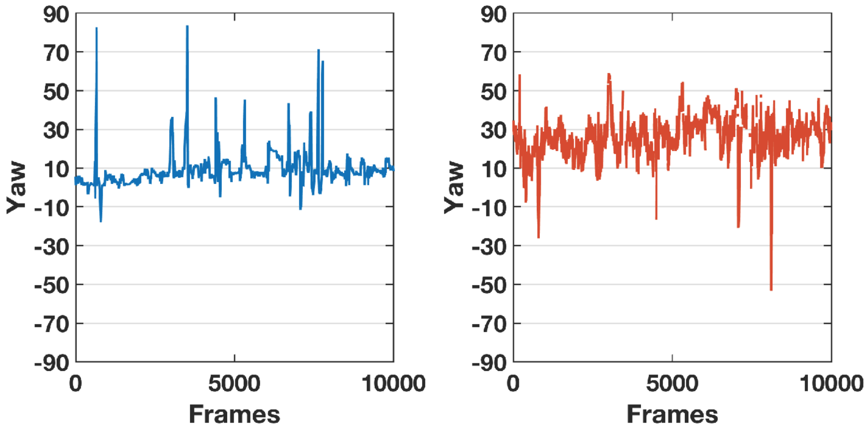
Fig. B.1 Histograms of the computed EAR after the extraction of the eye landmarks using Dlib in figures (a), (c) and (e), and OpenFace toolkit in figures (b), (d) and (f). The X-axis of OpenFace figures (b), (d), and (f) ranged to about 2.5, 33.85, and 13, respectively. However, the range was cut off to one for them to make clear visualisation of the histograms.

B.2 The Head Movements

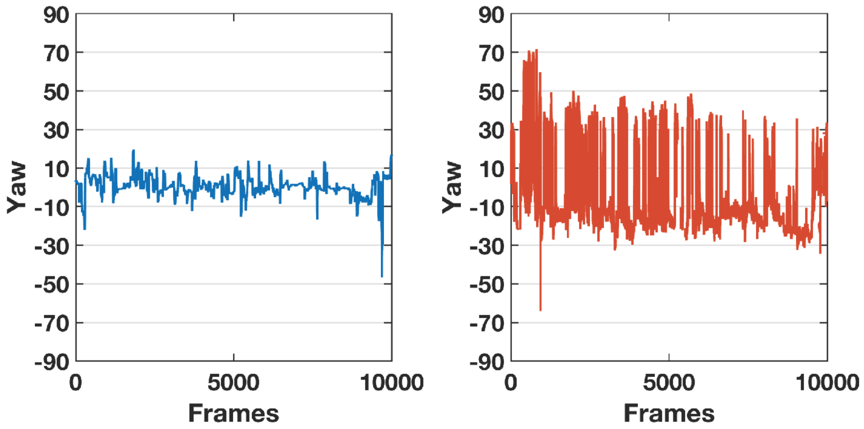
B.2.1 Examples of the Head Pose Estimation using Dlib and OpenFace



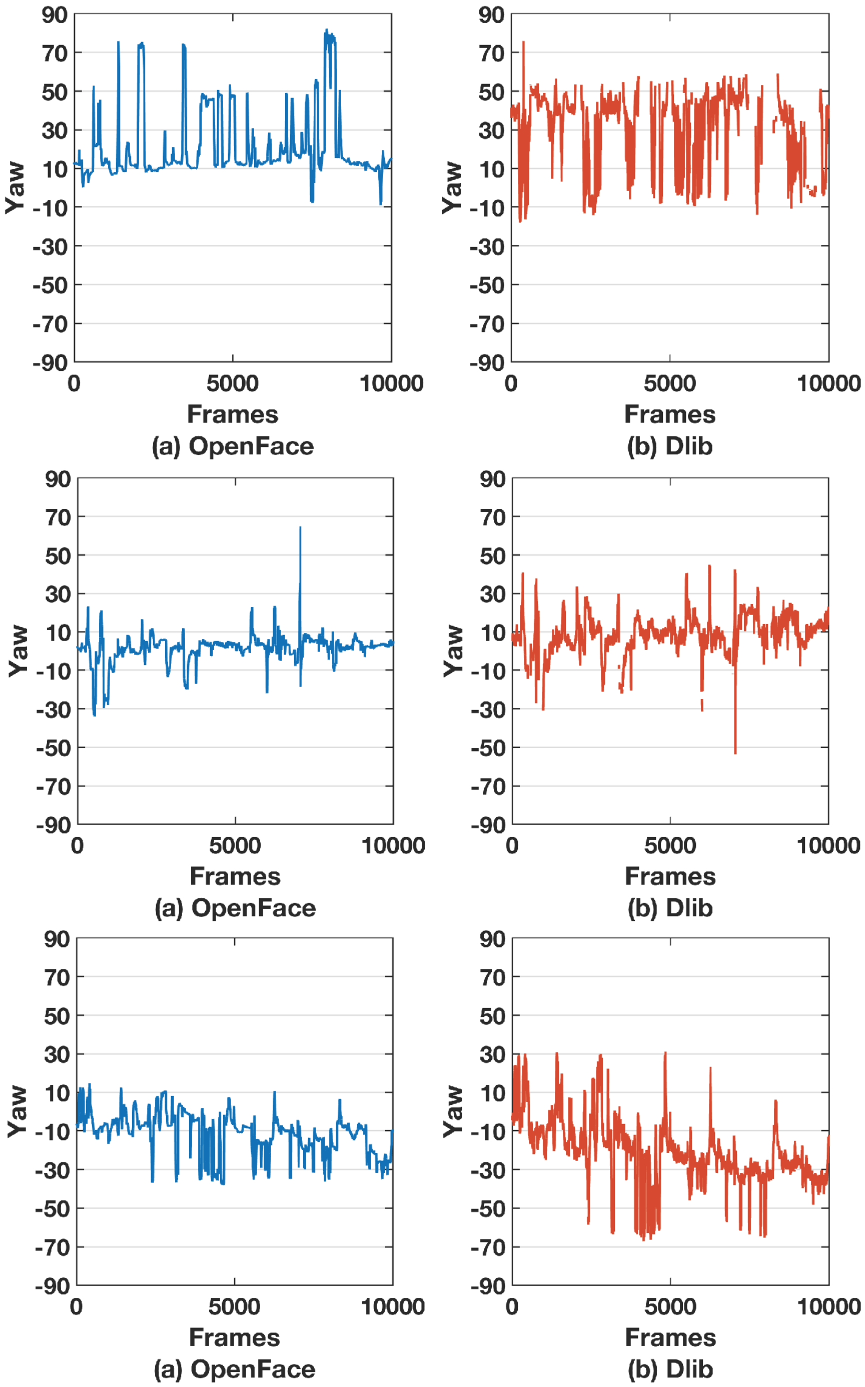
(a) OpenFace (b) Dlib



(a) OpenFace (b) Dlib



(a) OpenFace (b) Dlib



(a) OpenFace

(b) Dlib

(a) OpenFace

(b) Dlib

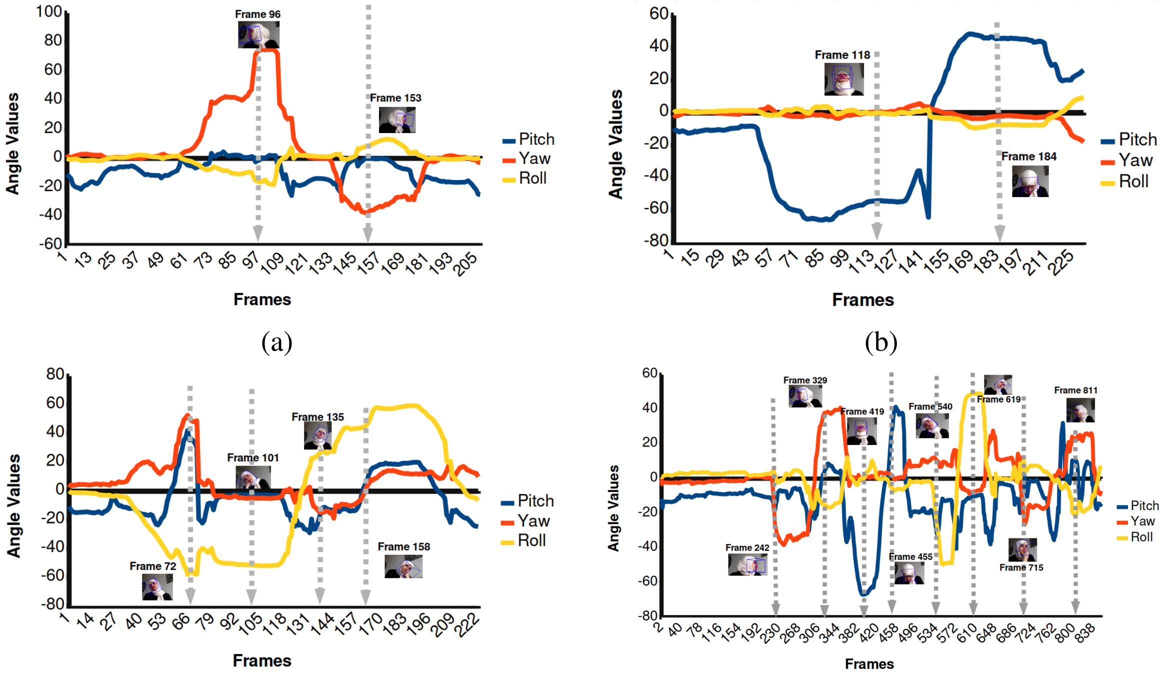
(a) OpenFace

(b) Dlib

Fig. B.2 The estimated yaw values of participants from different groups using both facial landmarks tracking techniques:OpenFace and Dlib.

B.2.2 Self-test for Measuring the Maximum and Minimum Angles to Capture by OpenFace

Figures B.3 and B.4 show the calculated three angles for different recordings at day and night times, respectively. It can be seen that when there is a high value, there is a small picture from the recording indicates the head orientation. Notably, the angle values are higher when the recording is made at night time. This may be affected by the low illumination in the room and rapid head movements that cause a dark side in the frame, which results in failure in the prediction of the angle value. From that, any value that is greater than$[-90, +90]$ is an outlier value, and that value is going to be linearly interpolated.



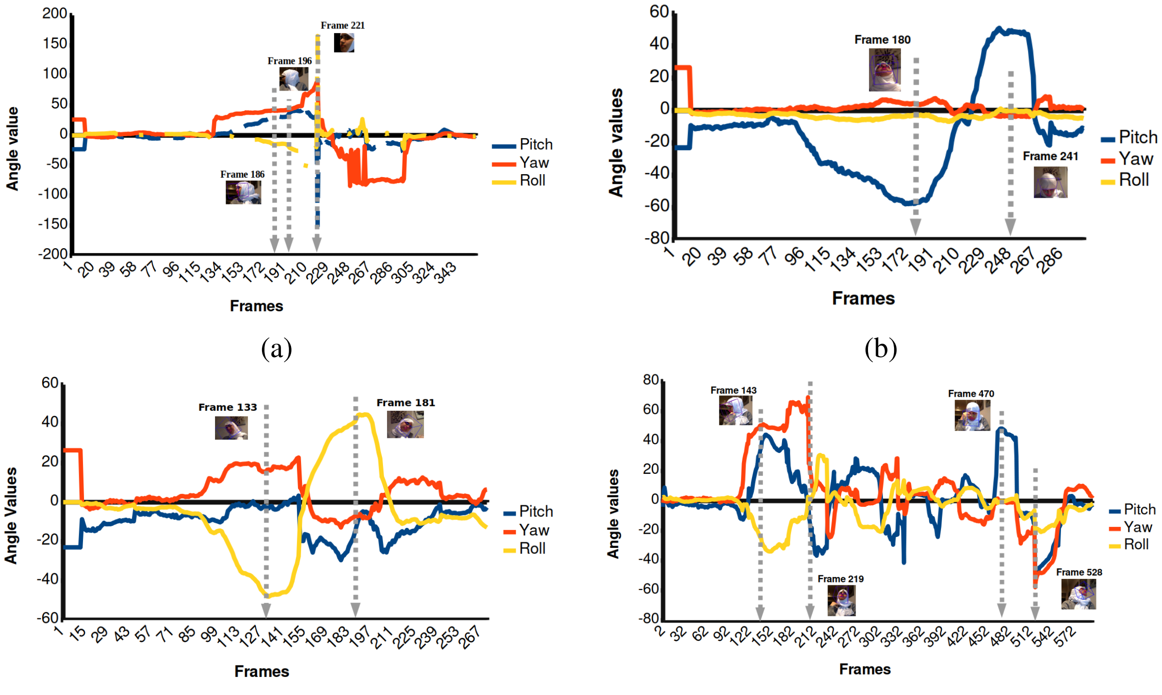
(a)

(b)

(c)

(d)

Fig. B.3 The calculated three angles for self-test videos at day time. The plots show several pictures describe the head orientation state when the angle value is high.

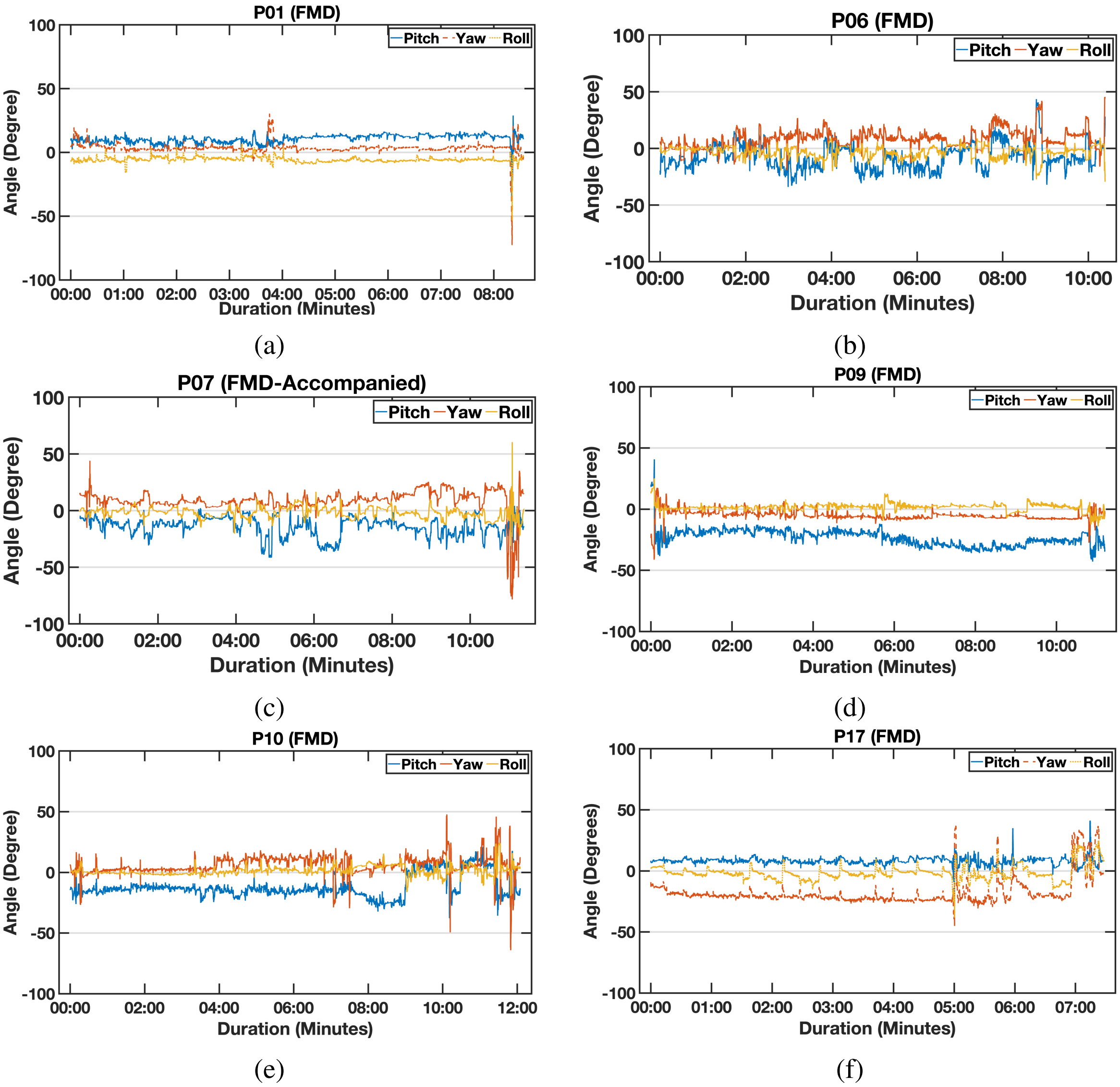


(a) (b)

(c) (d)

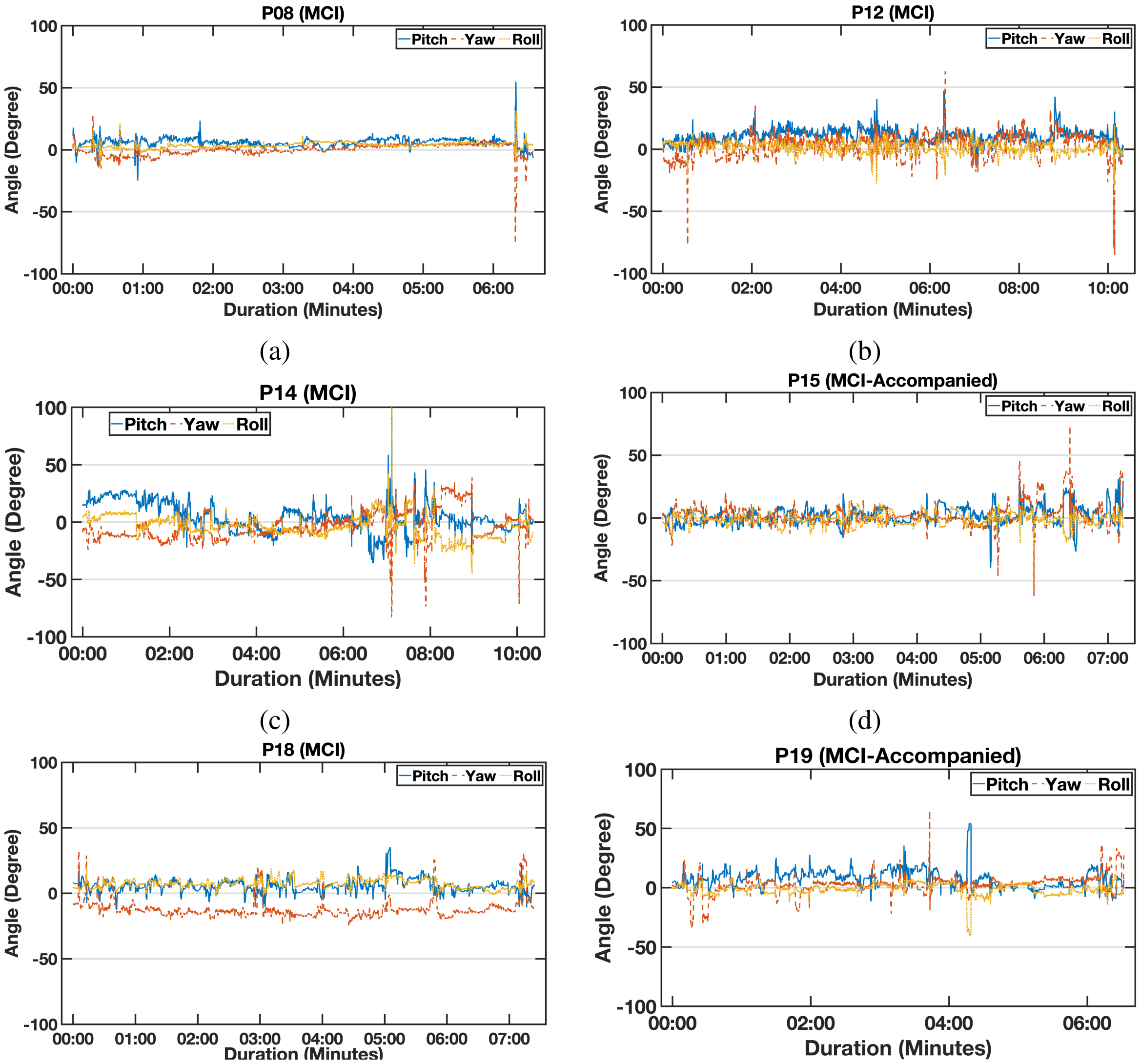
Fig. B.4 The calculated three angles for self-test videos in poor lighting conditions. The subfigures show several pictures describing the head orientation state when the angle value is high.

B.2.3 The Calculated Three Angles of Head Movements for the IVA_{18} Dataset



(a) (b) (c) (d) (e) (f)

Fig. B.5 The calculated three angles pitch, yaw, and roll values for each participant with FMD. The highlighted segments indicate the participant's turn to answer.

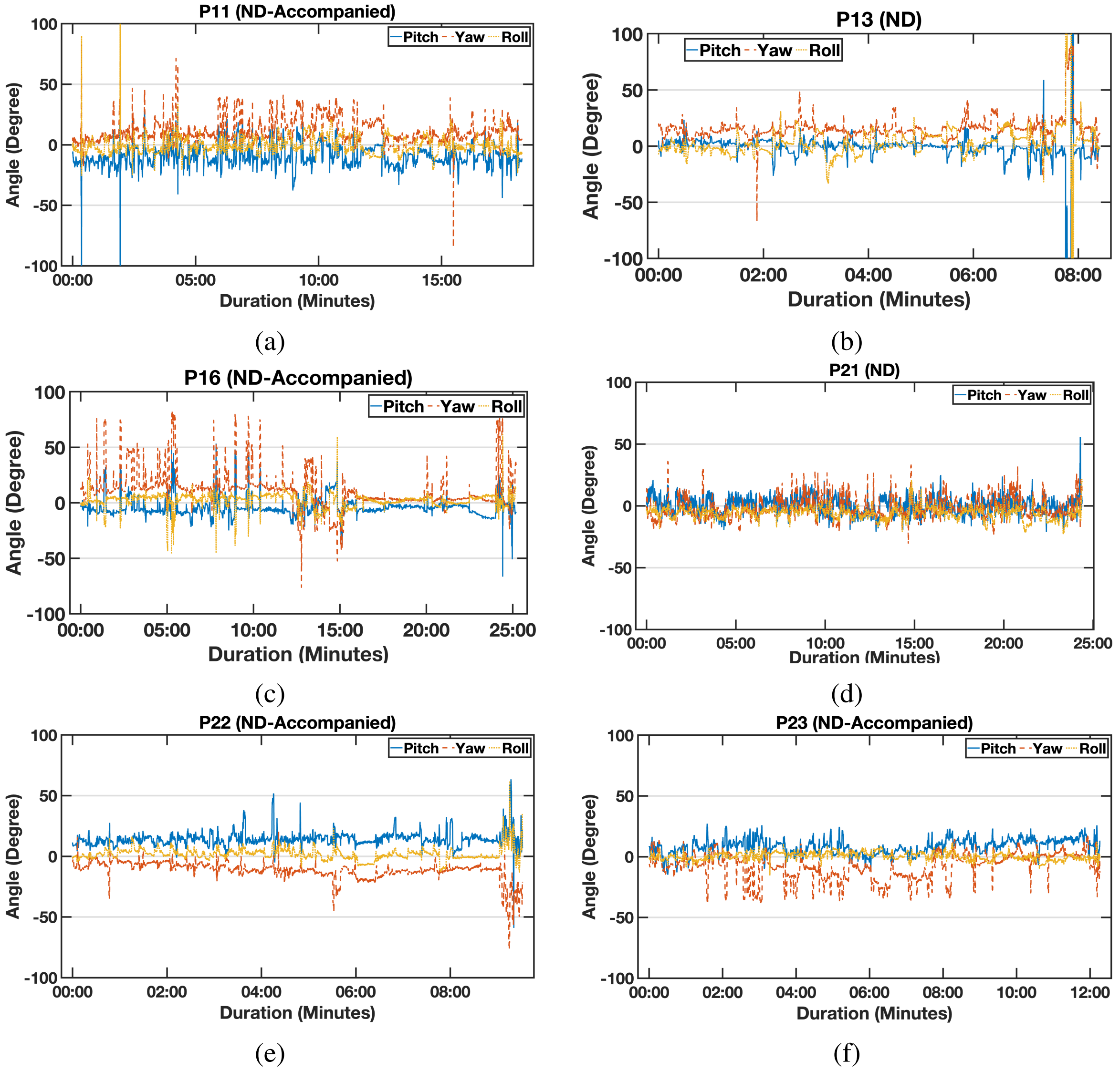


(a) (b)

(c) (d)

(e) (f)

Fig. B.6 The calculated three angles pitch, yaw, and roll values for each participant with MCI. The highlighted segments indicate the participant's turn to answer.



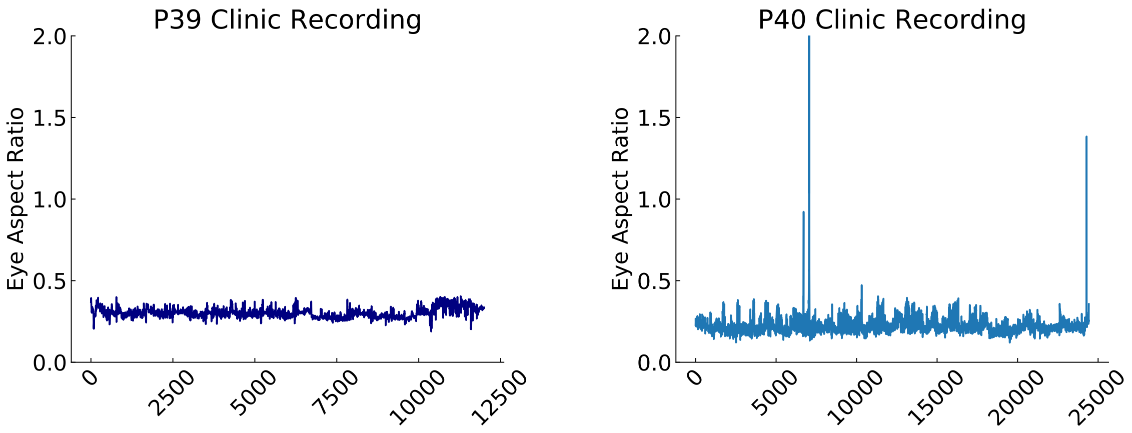
(a) (b) (c) (d) (e) (f)

Fig. B.7 The calculated three angles pitch, yaw, and roll values for each participant with ND. The highlighted segments indicate the participant's turn to answer.

Appendix C

The IVA_{34} Dataset

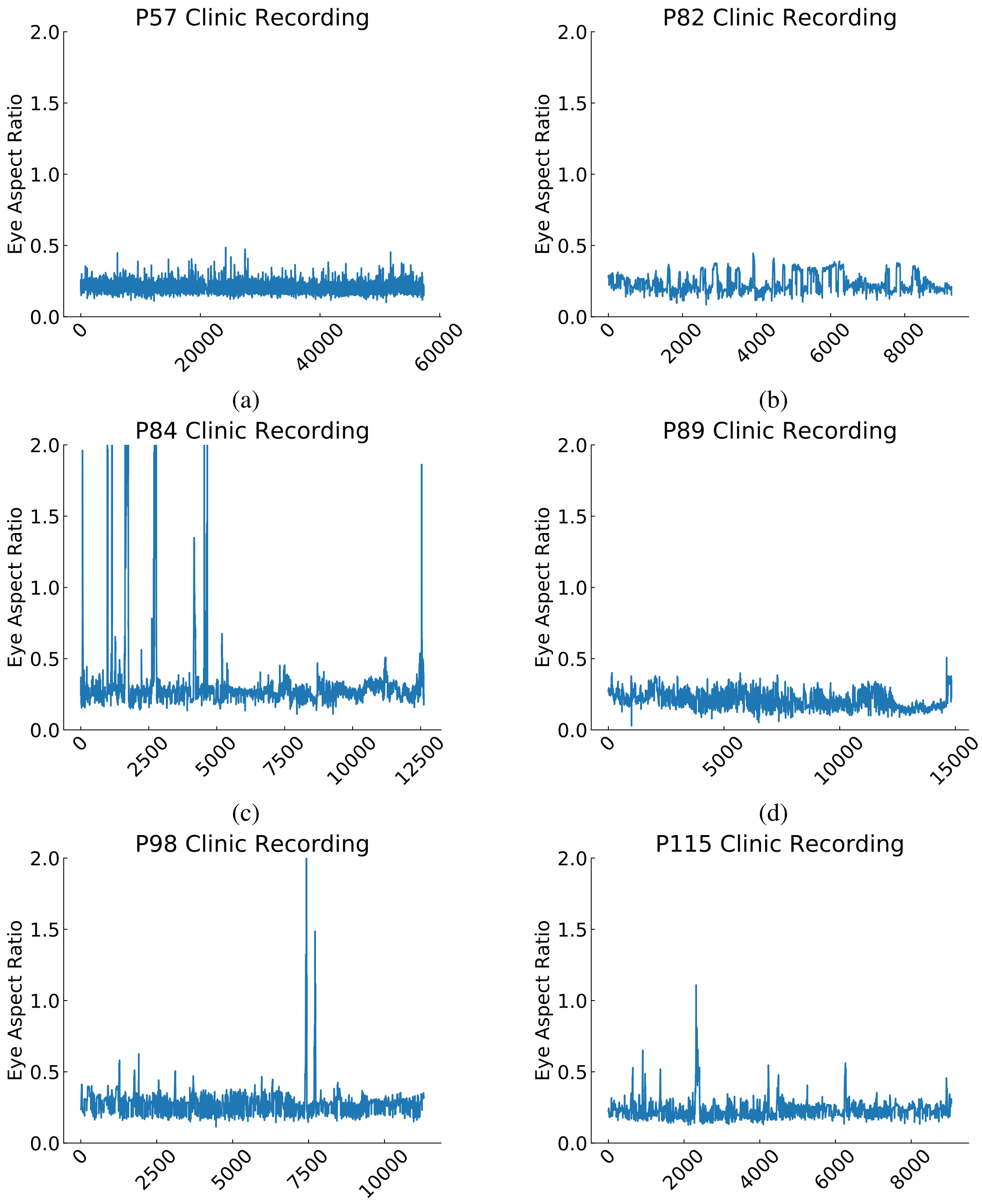
C.1 The Calculated EAR for the IVA_{34} Dataset



(a)

(b)

Fig. C.1 The calculated EAR values of different participants from data recorded at the **clinic** after removing the unnecessary data from the beginning and the end of the session (part 1).

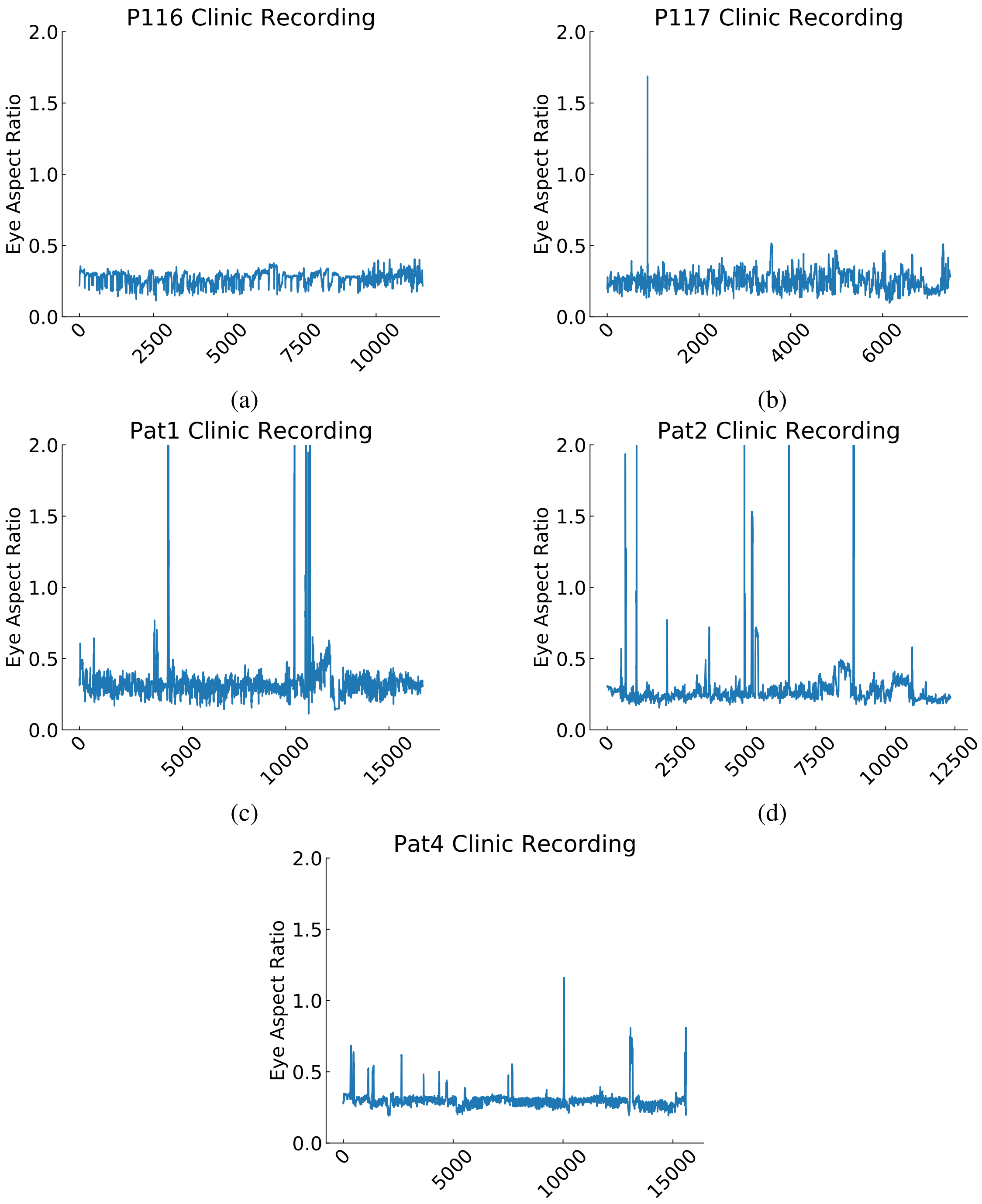


(a) (b)

(c) (d)

(e) (f)

Fig. C.2 The calculated EAR values of different participants from data recorded at the **clinic** after removing the unnecessary data from the beginning and the end of the session (part 2).



(a)

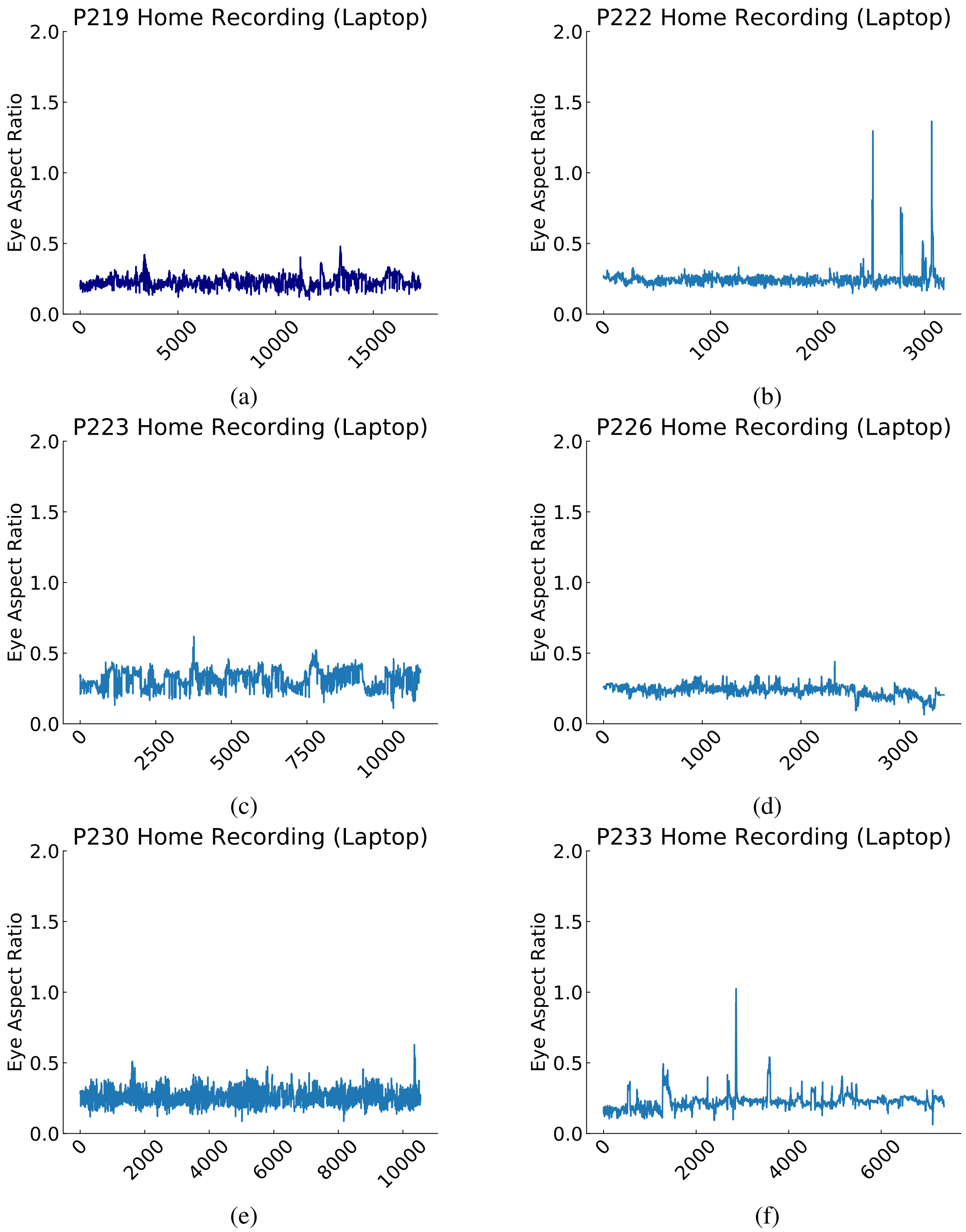
(b)

(c)

(d)

(e)

Fig. C.3 The calculated EAR values of different participants from data recorded at the **clinic** after removing the unnecessary data from the beginning and the end of the session (part 3).



(a)

(b)

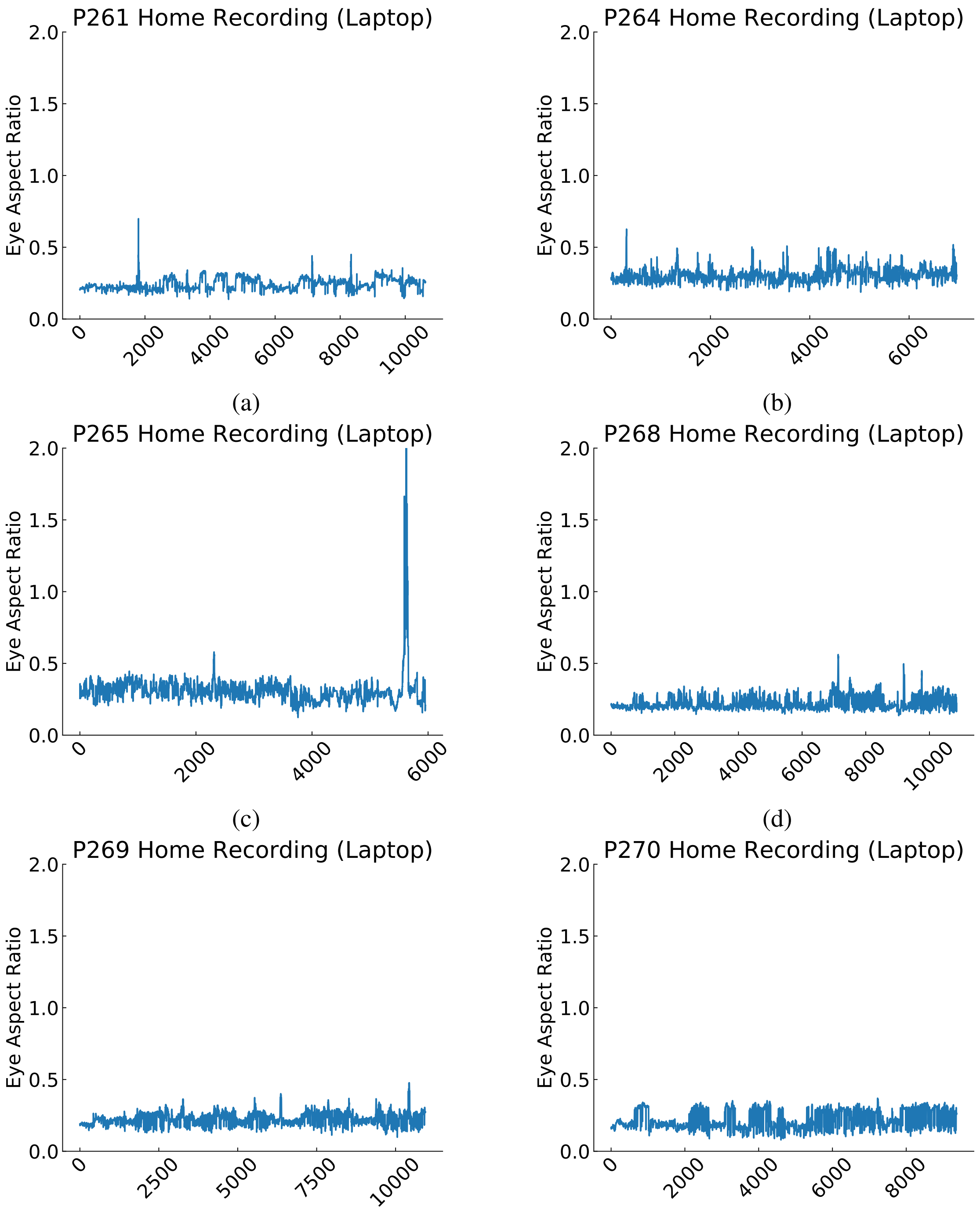
(c)

(d)

(e)

(f)

Fig. C.4 The calculated EAR values of different participants from data recorded at **home** after removing the unnecessary data from the beginning and the end of the session (part 1).



(a)

(b)

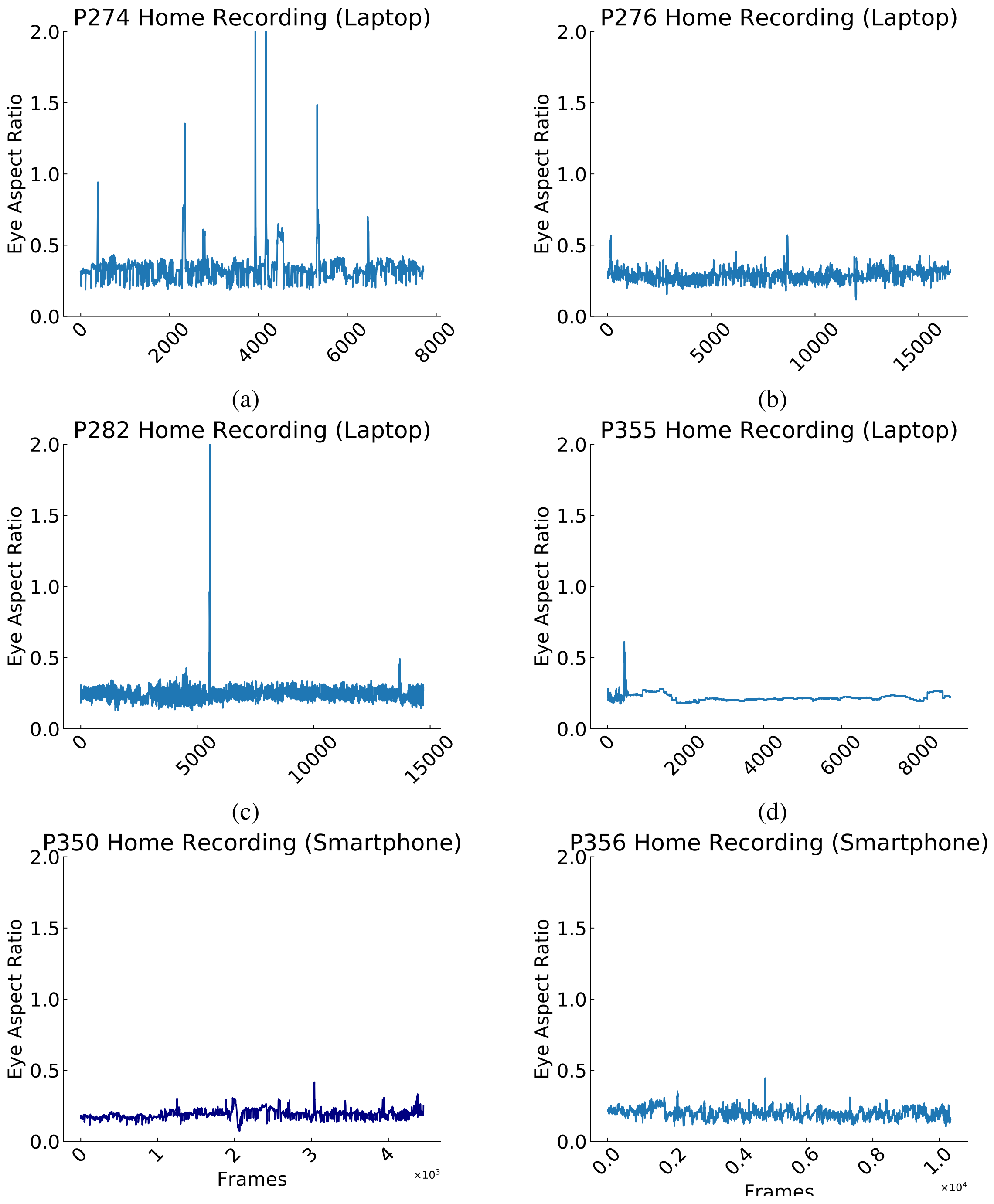
(c)

(d)

(e)

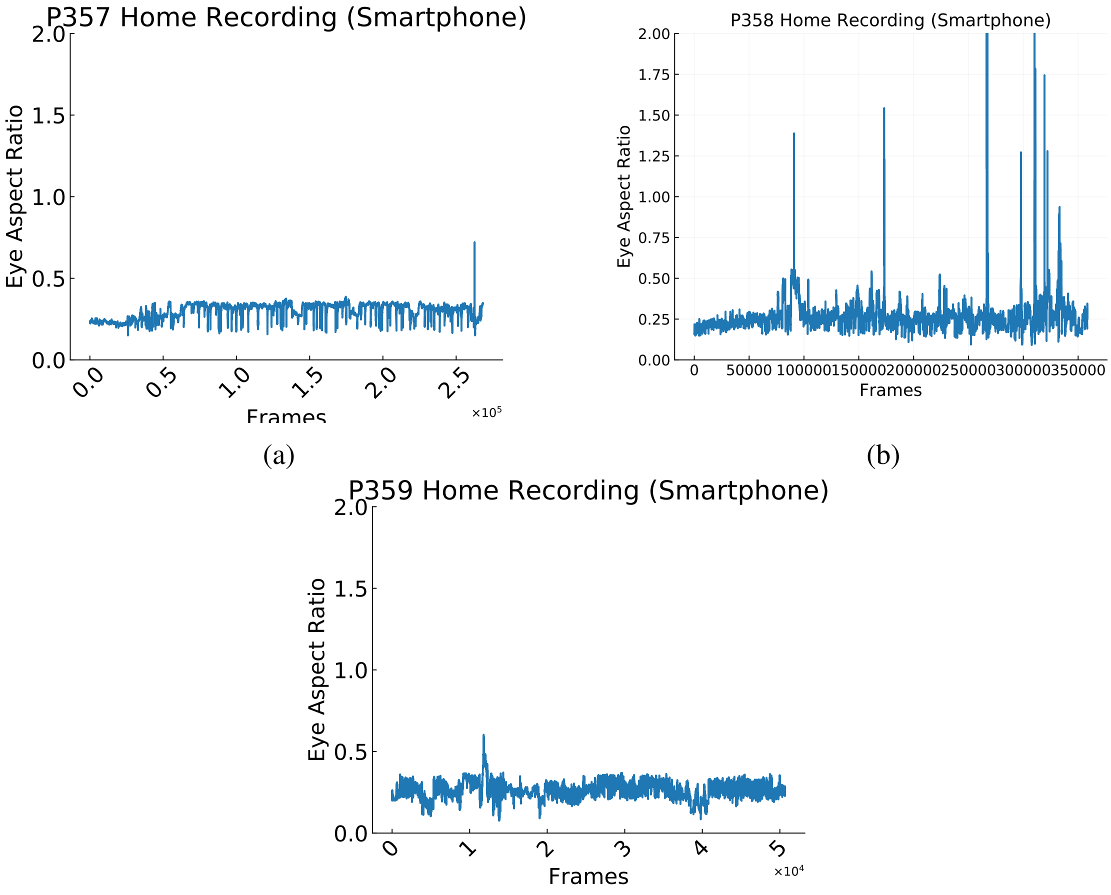
(f)

Fig. C.5 The calculated EAR values of different participants from data recorded at **home** after removing the unnecessary data from the beginning and the end of the session (part 2).



(a) (b) (c) (d) (e) (f)

Fig. C.6 The calculated EAR values of different participants from data recorded at **home** after removing the unnecessary data from the beginning and the end of the session (part 3).



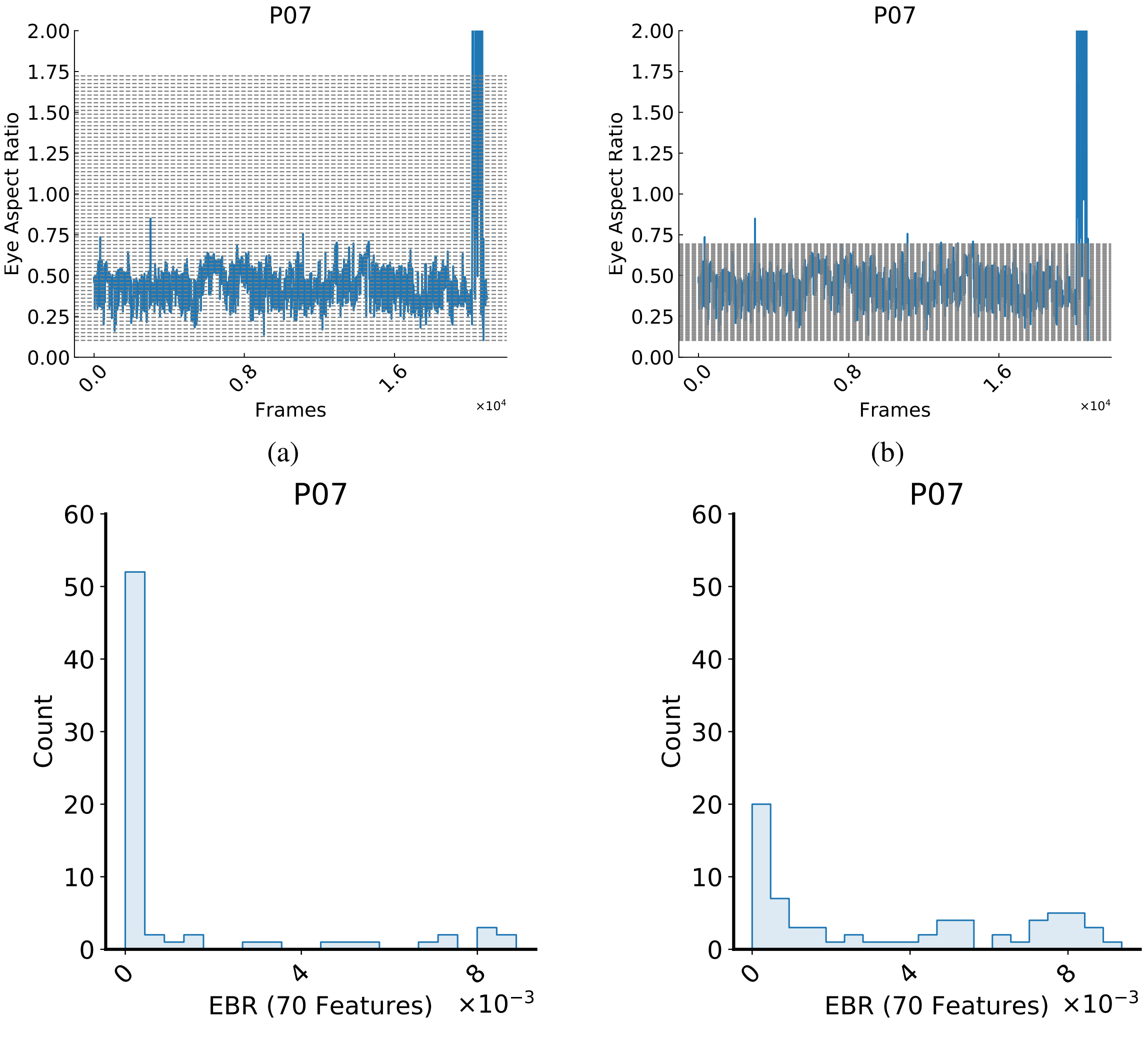
(a)

(b)

(c)

Fig. C.7 The calculated EAR values of different participants from data recorded at **home** after removing the unnecessary data from the beginning and the end of the session (part 4).

C.2 Thresholds Calculation for EAR using SD and IQR Approaches



(a) (b)

(c) (d)

Fig. C.9 The EAR with the calculated thresholds using the 3rd SD and IQR (1st row) and histogram figures show the EBR feature distribution (2nd row) for several participants in the IVA_{18} dataset (Part 1).

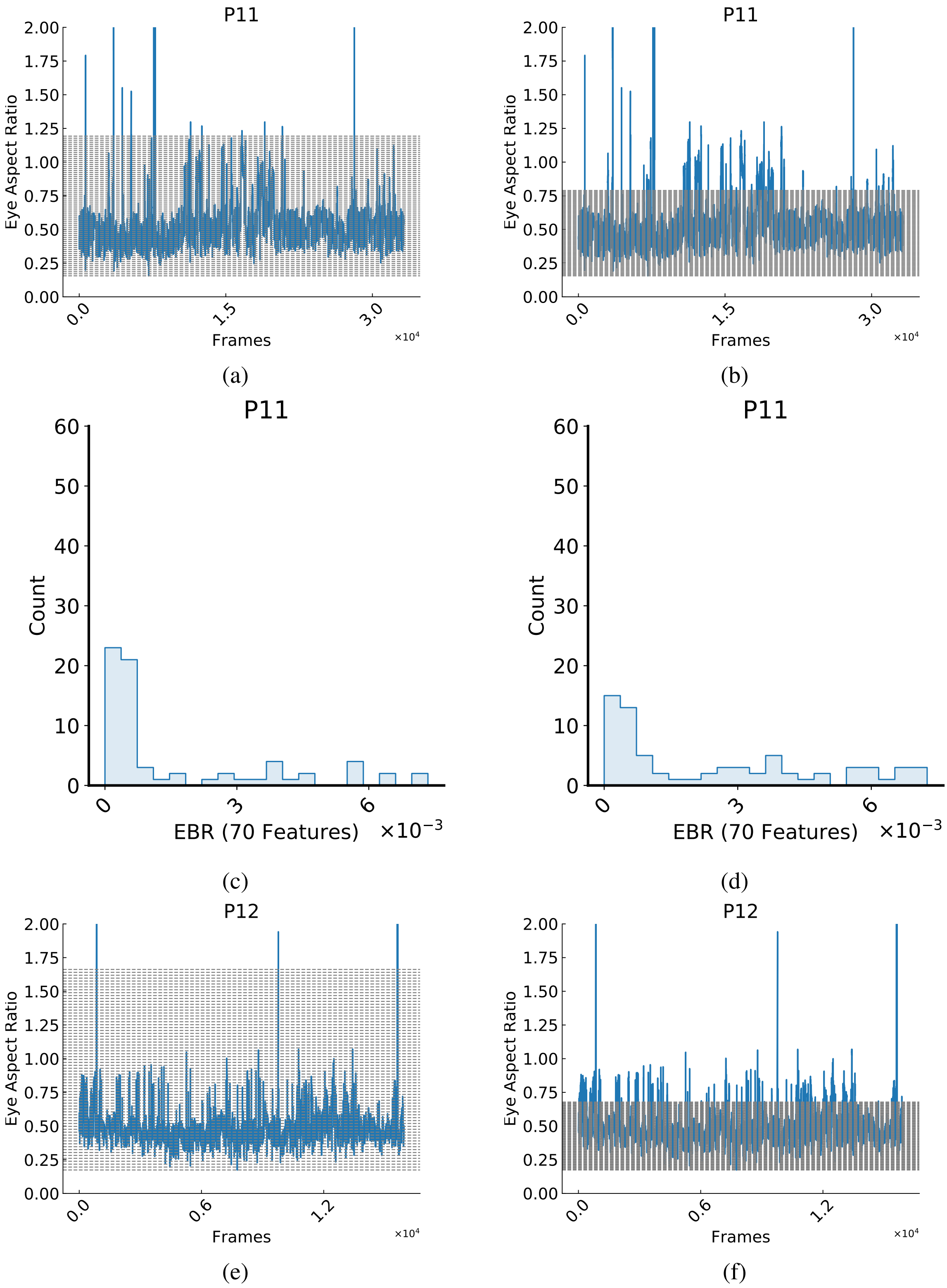
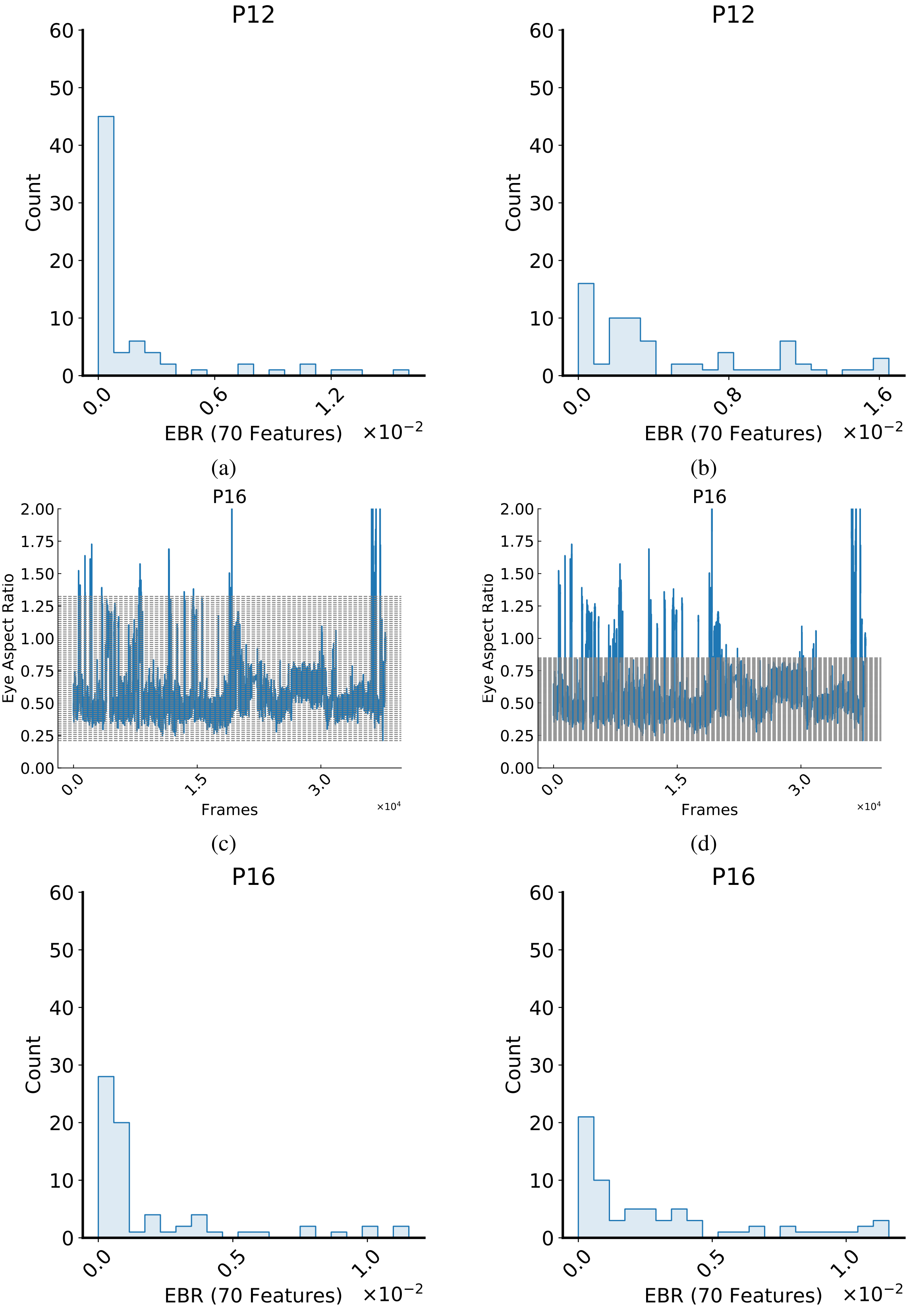


Fig. C.11 The EAR with the calculated thresholds using the 3rd SD and IQR (1st row) and histogram figures show the EBR feature distribution (2nd row) for several participants in the IVA_{18} dataset (Part 2).



(a)

(b)

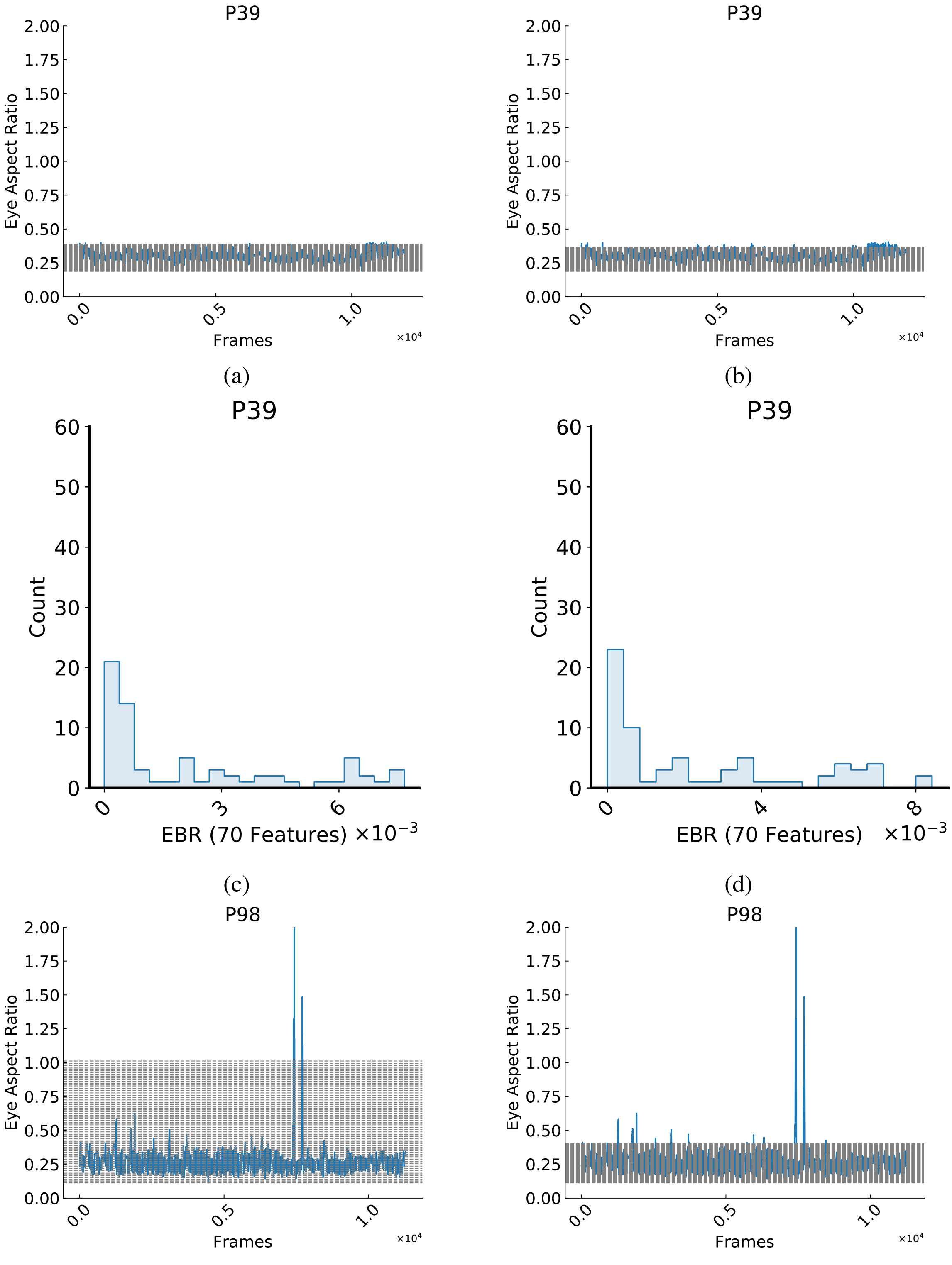
(c)

(d)

(e)

(f)

Fig. C.13 The EAR with the calculated thresholds using the 3rd SD and IQR (1st row) and histogram figures show the EBR feature distribution (2nd row) for several participants in the *IVA*$_{18}$ dataset (Part 3).



(a)

(b)

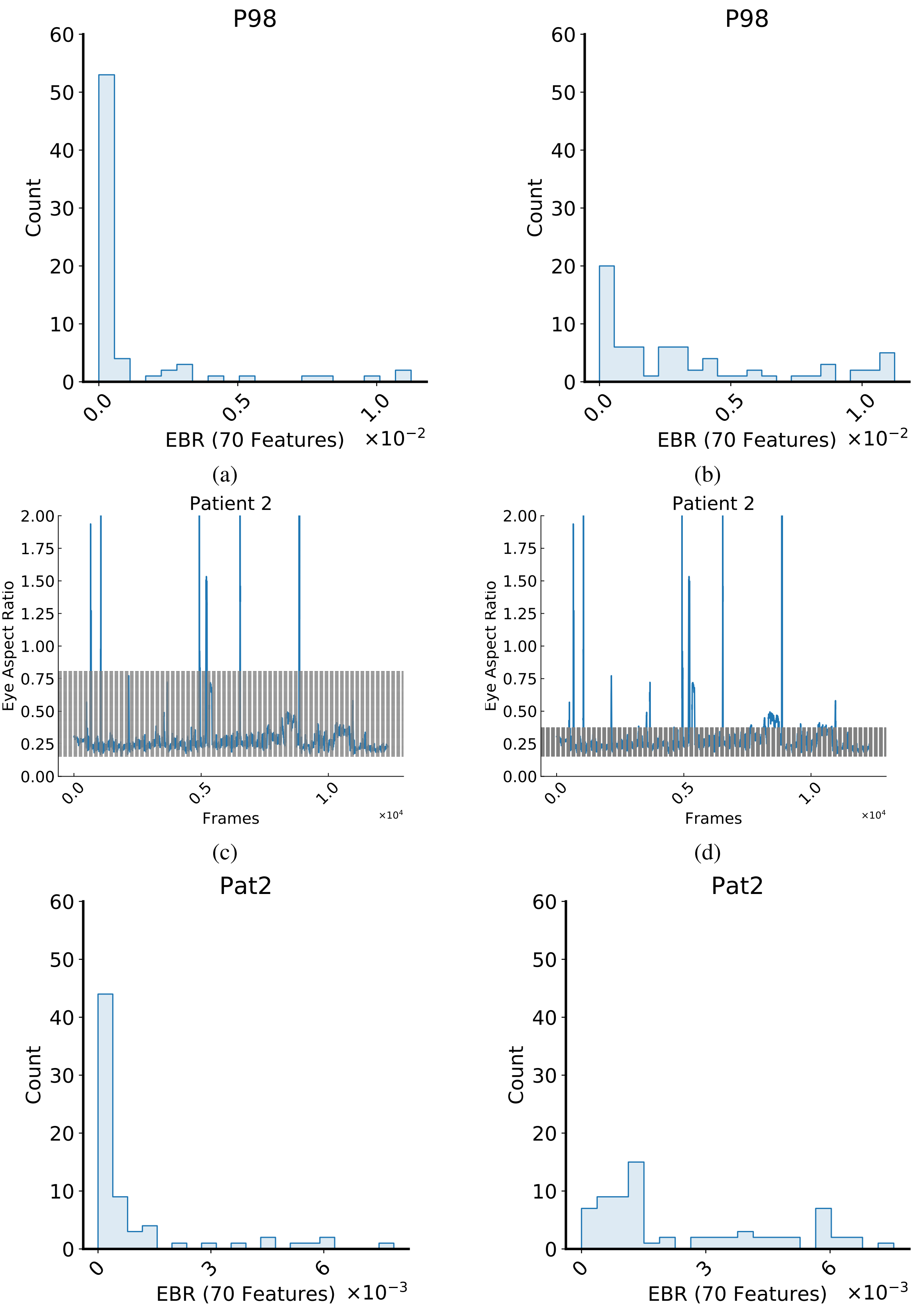
(c)

(d)

(e)

(f)

Fig. C.15 The EAR with the calculated thresholds using the 3rd SD and IQR (1st row) and histogram figures show the EBR feature distribution (2nd row) for several participants in the IVA_{34} dataset (Part 1).

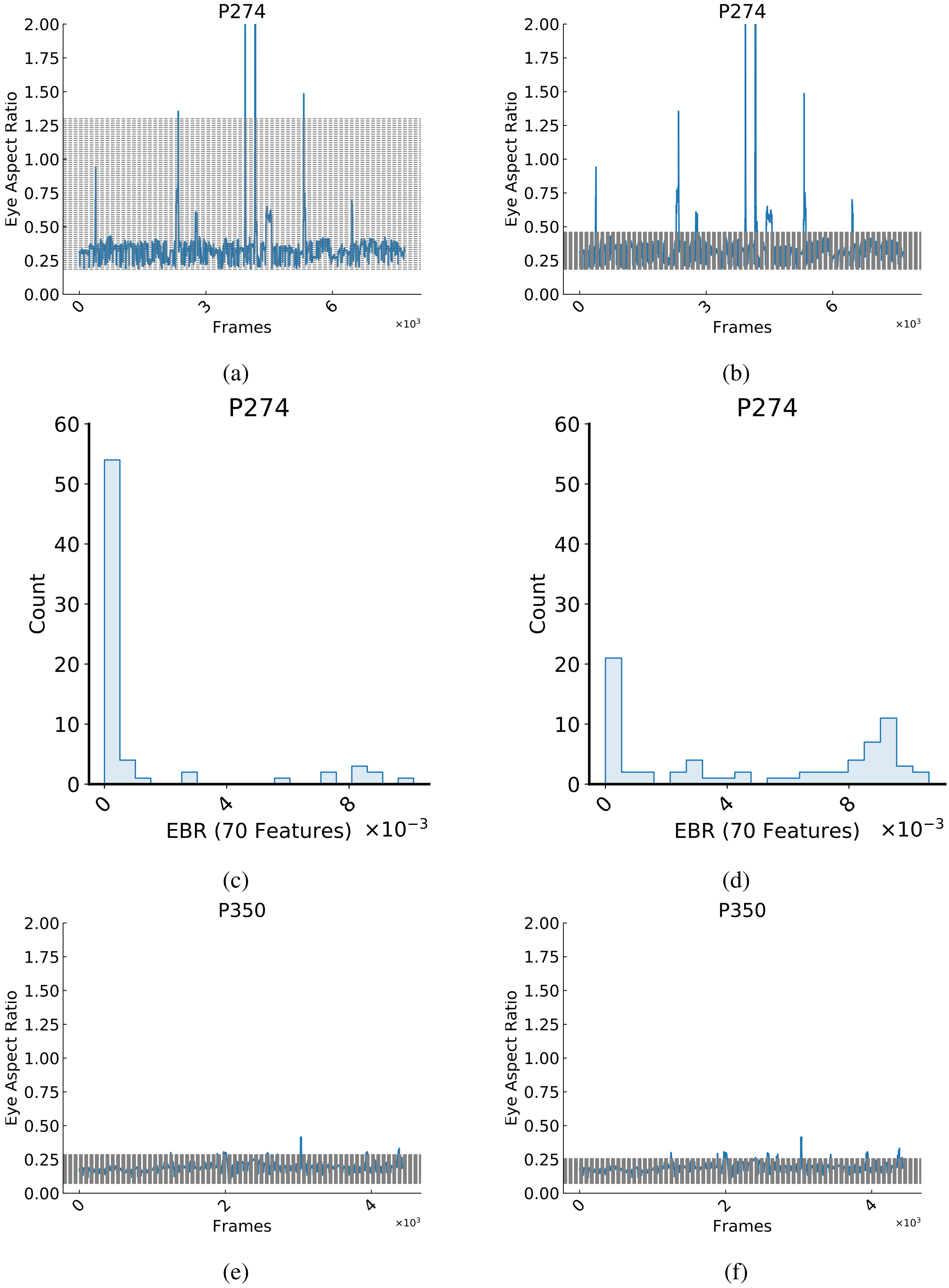


(a) (b)

(c) (d)

(e) (f)

Fig. C.17 The EAR with the calculated thresholds using the 3rd SD and IQR (1st row) and histogram figures show the EBR feature distribution (2nd row) for several participants in the IVA_{34} dataset (Part 2).



(a)

(b)

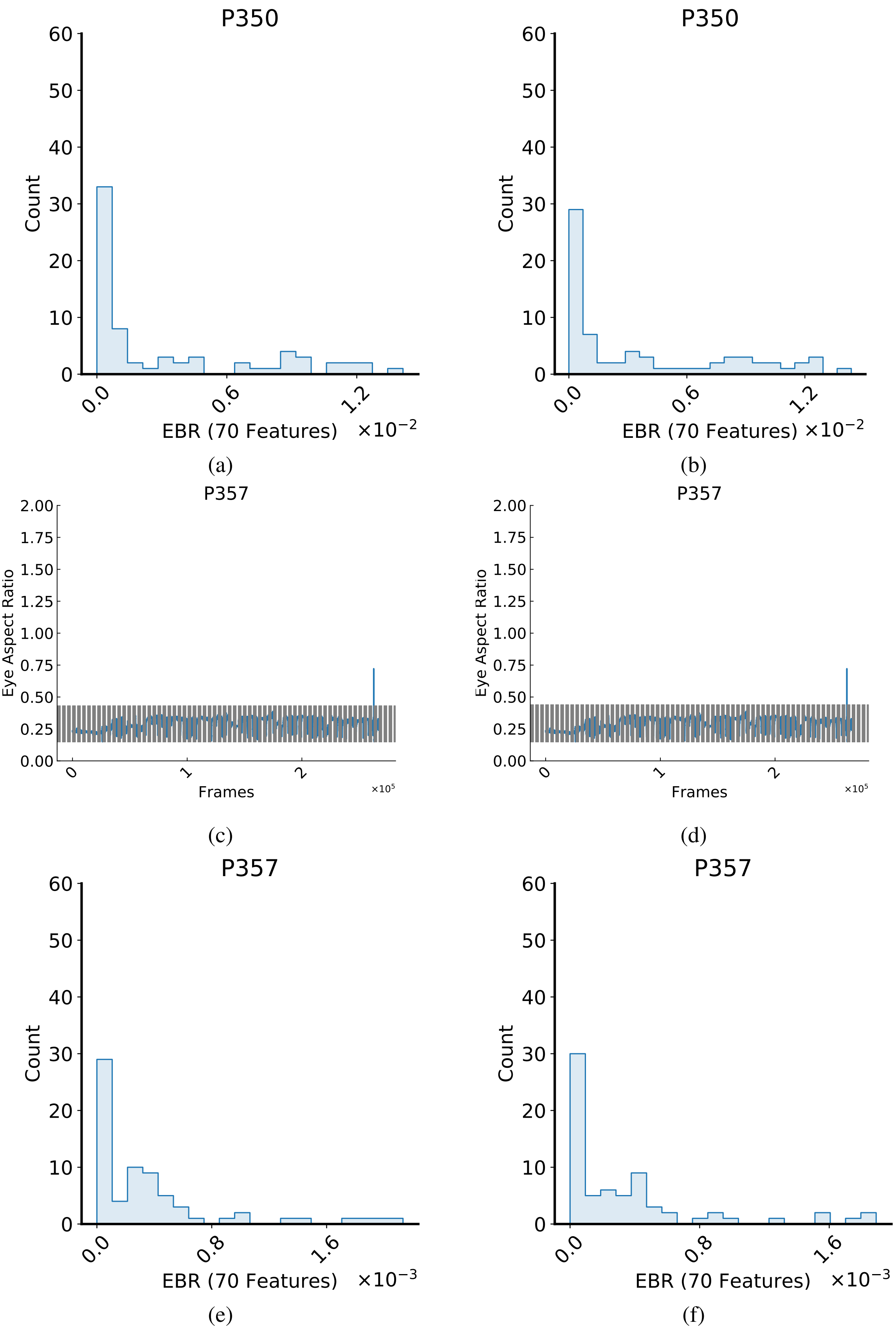
(c)

(d)

(e)

(f)

Fig. C.19 The EAR with the calculated thresholds using the 3rd SD and IQR (1st row) and histogram figures show the EBR feature distribution (2nd row) for several participants in the *IVA*$_{34}$ dataset (Part 3).

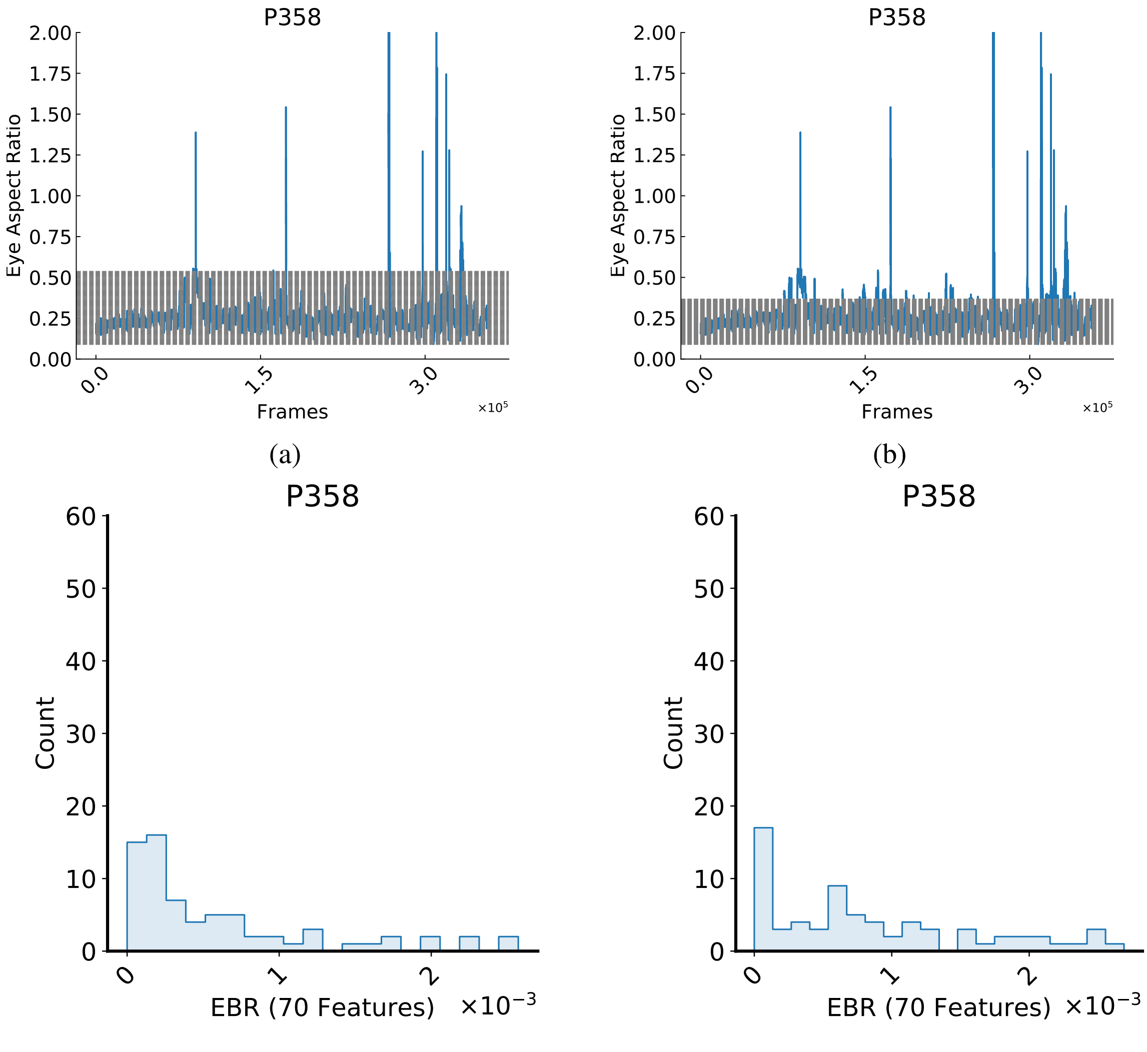


(a) (b)

(c) (d)

(e) (f)

Fig. C.21 The EAR with the calculated thresholds using the 3rd SD and IQR (1st row) and histogram figures show the EBR feature distribution (2nd row) for several participants in the *IVA*$_{34}$ dataset (Part 4).



(a)

(b)

(c)

(d)

Fig. C.23 The EAR with the calculated thresholds using the 3rd SD and IQR (1st row) and histogram figures show the EBR feature distribution (2nd row) for several participants in the IVA_{34} dataset (Part 5).