The
University
Of
Sheffield.

Audio Based Signal Processing and Computational Models for Early Detection and Prediction of Dementia and Mood Disorders

By

Sabah Al-Hameed

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

The University of Sheffield
Department of Electrical and Electronic Engineering

April 2020

Supervisors

Dr. Mohammed Benaissa
Dr. Heidi Christensen

Declaration

I confirm that this thesis is my work unless otherwise, I mentioned to studies done by other researchers in the text. The contents of this project are original and not presented to other universities, or awards in any multiversities. However, most of the experiments in this study have been published as a journal and conference papers.

SABAH Al-Hameed
April 2020

Acknowledgements

First, I would like to thank god who gave me the strength and determination to complete this journey, I'm also grateful for all of his blessings.

Special thanks to my supervisor Dr Mohammed Benaissa who support and guide me to complete and publish my work. Also, I would like to thank Dr Heidi Christensen for her encouragement and knowledge in this field which helped me a lot to understand many things and polish my work to the level where a researcher should be.

I'm grateful for my uncle Mr Moshreq Kadhim who stands on my behalf to overcome a serious problem which could prevent me from accomplishing this journey.

To my mom, grandmom, who I will always be grateful and because of their support and kindness, I am what I am right now.

To my wife and my children: Ali, Fatimah and Hussien, thanks for your continuous support and trust. I will always be grateful for your kindness.

Abstract

Neurodegenerative diseases causing dementia are known to affect a person's speech and language. There is an increasing emphasis on earlier diagnosis of neurodegenerative disorders as evolving treatments are likely to be more effective before irreversible changes have occurred in the brain. The incorporation of novel methods based on the automatic analysis of speech signals may provide more information about a person's ability to interact, which could contribute to the diagnostic process.

This thesis demonstrates that purely acoustic features, extracted from recordings of patients' answers to a neurologist's questions in a specialist memory clinic can support the initial distinction between patients presenting with cognitive concerns attributable to progressive neurodegenerative disorders (ND) or Functional Memory Disorder (FMD, i.e., subjective memory concerns unassociated with objective cognitive deficits or a risk of progression). The thesis also shows that the acoustic features extracted from speech recordings for patients describing a picture can be used to construct a non-invasive and simple tool to infer early signs of dementia of Alzheimer's Disease (AD). This is further developed to firstly, identify patients with mild cognitive impairments and secondly to show a capability to assist the doctors in monitoring the progression of AD by predicting the MMSE scores in a longitudinal dataset.

Further, novel acoustic features are introduced in this thesis that correlate with mood disorders such as major depression and bipolar. Combing the newly extracted features with state of the art features, led to developing a language-agnostic screening system for depression and bipolar disease. Finally, the results obtained show the discriminative power of purely acoustic approaches that could be integrated into diagnostic pathways for patients presenting with memory concerns or mood disorders. These approaches are computationally less demanding than methods focusing on linguistic elements of speech and language that require automatic speech recognition and understanding.

Contents

List of Figures

List of Tables

List of Publications

Journal papers

- Al-Hameed S, Benaissa M, Christensen H, Mirheidari B, Blackburn D, et al. (2019) A new diagnostic approach for the identification of patients with neurodegenerative cognitive complaints. PLOS ONE 14(5): e0217388.https://doi.org/10.1371/journal.pone.0217388.

Conference Papers

- Al-Hameed S, Benaissa M, Christensen H. Simple and robust audio-based detection of biomarkers for Alzheimer's disease. In7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT) 2016 (pp. 32-36).
- Al-Hameed S, Benaissa M, Christensen H. Detecting and predicting alzheimer's disease severity in longitudinal acoustic data. InProceedings of the International Conference on Bioinformatics Research and Applications 2017 2017 Dec 8 (pp. 57-61). ACM.

In preparation Journals

- Al-Hameed S, Benaissa M, Christensen H. Towards the language-agnostic depression assessment tool.
- Al-Hameed S, Benaissa M, Christensen H. Predicting Bipolar states using acoustic characteristics.

List of Abbreviations

ACE	Addenbrooke’s Cognitive Examination
AD	Alzhiemer’s Disease
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
AUC	Area Under Receiver Operating
BDI	Beck DepressionIndex
BDNF	Brain Derived Neurotropic Factor
BNT	Boston Naming Test
bvFTD	behavioural variant frontotemporal dementia
CSF	Cerebro Spinal Fluid
CT	Com-puted Tomography
DCNN	Deep Convolutional Neural Network
DF	Degree of Freedom
DLB	Dementia with Lewy bodies
DPD	Pseudo-Dementia
DSM	Statistical Manual of Mental Disorders

EEG	Electroencephalogram
EER	Equal Error Rate
FFT	Fast Fourier Transform
FLD	Frontal Lobe Dementia
FMD	Functional Memory Disorder
FNR	False negative rate
FN	False Negatives
FPR	False positive rate
FP	False Positives
FS	Feature Selection
GABA	Gamma Amino Butyric Acid
GAD-7	Generalized Anxiety Disorder 7
GMM-UBM	Gaussian Mixture with the Universal Background
GMM	Gaussian Mixture Model
HAMD	Hamilton Rating Scale for Depression
HC	Healthy Control
HNR	Harmonic to Noise Ratio
LDA	Linear Discriminant Analysis
MAE	Mean Absolute Error
MARSD	Montgomery – Åsberg Depression Rating Scale
MCI	Mild Cognitive Impairment

MFCC	Mel Frequency Cepstral Coefficients
MHH	Motion History Histogram
MMSE	Mini-Mental State Examination
MoCA	Montreal Cognitive Assessment
MRI	agnetic Res-onance Imaging
ND	Neurodegenerative Diseases
nfPPA	Progressive non-fluent Aphasia
NHR	Noise to Harmonic Ratio
PET	Positron EmissionTomography
PHQ-9	Patient Health Questionnaire-9
PPV	Positive predictive value
QIDS	Quick Inventory of Depressive Symptomatology
RBF	Radial basis Function
ReLU	Rectified Linear Unit
RFE	Recursive Feature Elimination
RMSE	Root Mean Square Error
ROC	Receiver Operating Characterstic
SAS	Zung Self-Rating Anxiety Scale
SCD	Subjective Cognitive Decline
SDS	Zung Self-Rating Depression Scale
SD	Semantic Dementia

SE	Spectral Entropy
SGD	Stochastic Gradient Descent
SMOT	Minority Over Sampling Technique
SMO	Sequential Multiple Optimisation
SPECT	Single Photon Emission Computed Tomography
SRO	Spectral Roll-Off
SSC	Spectral Subband Centroids
SS	Spectral Spread
STD	Standard Deviation
SVM	Support Vector Machine
SVR	Support Vector Regression
TNR	True Negative Rate
TN	True Negatives
TPR	True Positive Rate
TP	True Positives
UAR	Unweighted Average Recall
VD	Vascular Dementia
WAIS	Wechsler Adult Intelligence Scale
WMS	Wechsler Memory Scale
WVR	word Vector Representations
ZCR	Zero Crossing Rate

Chapter 1

Introduction

Dementia and depression are the two most common causes of mental disorders affecting human life. The onset of Alzheimer's dementia (AD) and the development of depression share similar physical diseases and environmental factors. Undefined illnesses such as general fatigue and headache are recognised as warning symptoms in the very early stage of AD, these symptoms are shared with the pathology of depression. It is believed that symptoms of depression before the loss of cognition serve as a risk factor to the development of AD [3].

Dementia is a general term used to describe a set of symptoms that may include loss or decline in memory and other cognitive abilities. Although dementia is regularly spotted in older people, it is not an outcome of ageing [4]. Dementia starts typically with memory problems but also may involve difficulties in planning, performing daily tasks, struggle in communication and changes in personality, mood, and judgement [5]. It is essential to mention that sometimes, depression can be confused with the early signs of dementia, and a considerable number of people with dementia are also depressed [6]. However, dementia symptoms worsen over time and as the disease progresses to the final stage, people with dementia lose the ability to speak and even forget to eat and drink. As a result, demented people require 24-hour care and monitoring. In recent years, the number of people living with dementia has increased significantly. In 2015, the Alzheimer association organisation published a report for Alzheimer's disease facts and figures for the United States. The report revealed two majors public health issues; first, it indicated that between 2000 and 2013, there was a rapid increase of 71% in death statistics among the elderly caused by AD disease. Secondly, the cost estimate in 2014 for dementia

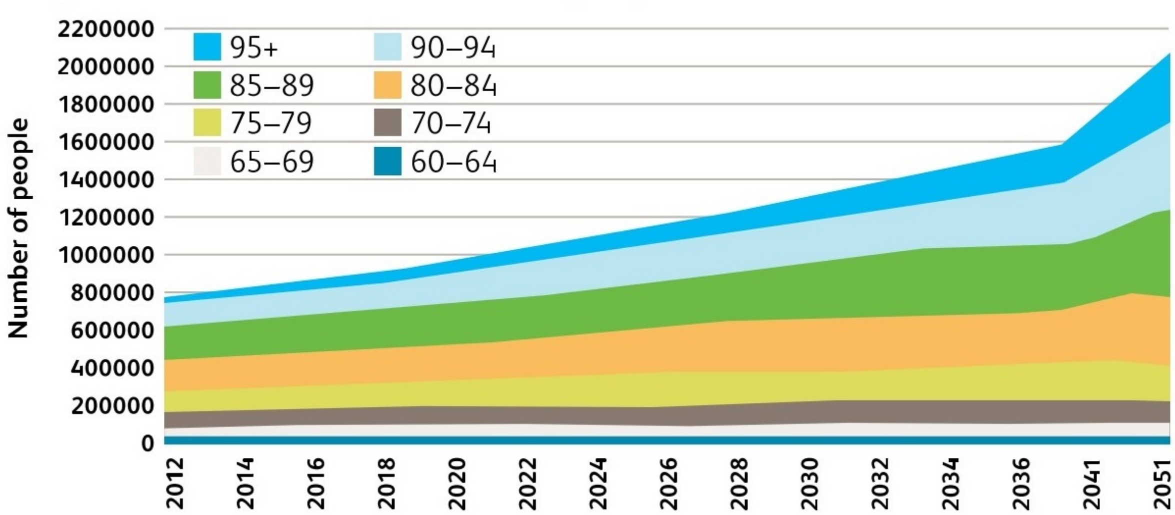


Figure 1.1 Projected prevalence of dementia in the UK up to 2051 adopted from [1]

diseases care is more than $217 billion per year. Also, the report stated that there are 5.3 million Americans having dementia, and this number was predicted to triple by 2050 with a massive impact on the government economy [7]. While in the UK, there are about 850,000 patients suffering from dementia with overall care cost exceeding £26.3 billion per year, and the number of patients is expected to grow to more than 2 million by 2051 [8] see Fig 1.1. Furthermore, the dementia prevalence around the world estimated at 44.3 million people with a cost at $604 billion per year, and this number is predicted to reach 152 million within the next 30 years [9]. The two common causes of dementia are Alzheimer's Disease (AD) (represents over 60 % of all cases) and Vascular Dementia (VD) (20%) [10, 11].

Depression ranges in severity from mild, short episodes of sadness to severe, persistent depression [12]. Clinical depression is a psychiatric mood disorder, resulting from a sudden stressful event affecting an individual's life. This leads to a continuous feeling of sadness, negativity, and makes it hard to handle everyday responsibilities. According to the World Health Organisation [13], depression is the leading cause of disability worldwide, and currently, there are more than 300 million people of different ages suffering from depression. In the UK, there are 2.69 million people (4.5% of the total population) diagnosed with depressive disorder [14]. The annual estimation cost for depression in the UK ranges between £7-£9 billion and that includes expenditures of preventive treatment, medical treatment, rehabilitation and care measures [15, 16]. While in the United States, the number of depressed people reported at 17.49

million (5.9% of the total population) [14], with total depression treatment cost around $71 billion [17]. Depression often triggers suicidal thoughts to end one's life [18, 19], and over 800,000 people worldwide die every year from committing suicide [20]. It has been confirmed that at least 50% of those who committed a suicide validate the condition of clinical diagnosis of depressive disorders [21, 22]. Extreme depression affecting not only the brain but also the heart, and therefore increases the risk for a range of medical conditions such as; cardiovascular disease, AD, vascular dementia, cancer, and stroke [19, 23, 24].

Speech and language characteristics are influenced in both disorders; early dementia affects para-linguistic and prosodic features (for example pitch, loudness, duration, silence, hesitations, and spectral aspects); furthermore, demented people start to forget words (loss of vocabulary) increase the use of filler words and at later stages of dementia, semantics impoverishment and other language defects can clearly be identify [4, 25, 26]. Depression, is characterised with a variety of clinical profiles; well-known symptoms include functional impairments, continuous low mood or sadness and moving or speaking more slowly than usual [27]; thus speech features that characterise mood, emotional state, and voice quality of the depressed person are all affected.

1.1 Motivation

The lack of a cure for dementia demands an early detection of the condition. The existing treatments are effective, particularly in the prior stages of the condition, especially before the occurrence of the irreversible changes in the brain. Hence, earlier diagnosis can improve quality of life and slow down the progress of the disease.

The early identification of patients with neurodegenerative disorders is a challenging task due to a lack of accurate predictive bio-markers suitable for routine screening or stratification. Bio-markers capable of identifying patients at high risk of developing the commonest cause of progressive cognitive decline, AD, pre-symptomatically exist but are either expensive and only available in very few centers (e.g. amyloid Positron Emission Tomography) that expose people to radiation or are invasive (e.g. amyloid and tau testing in the cerebrospinal fluid) and not suitable for for screening at the interface between primary and specialist care patients.

Tools for screening for AD exist, but currently do not work sufficiently well. For instance, the Dementia-Detection (DemTect) or the Montreal Cognitive Assessment (MoCA) are a one-page screening tools that can be administered by a trained examiner in 10 minutes. However, both produce learning effects which limit the number of possible administrations [28]. In addition, current cut-offs have poor specificity and have not been tested in people with functional memory disorder (who general practitioners may consider referring to specialist services because of their cognitive complaints) [29].

Concurrently, depression diagnosis, and treatment is determined on the subjective assessment of a wide range of profiles that characterise multiple endophenotypes. Although there were several biological markers associated with depression, for example, genetic abnormalities [30], low serotonin levels [31] and neurotransmitter dysfunction [32], to date, no specific biomarker has been identified. Therefore, the lack of an objective criterion hampers the efforts of clinical services for either depression or suicidality. This risks optimal patient care, twisting an already high burden on health, social, and economical utilities.

The current diagnostic tool for both depression and suicidality is based on assessments by interviews, for example, Hamilton Rating Scale for Depression and Suicide Probability Scales. These methods measuring the severity of symptoms and behaviours observed in both conditions quantified by the test scores. Using this method of assessment is not straightforward; the given scores are sensitive to the patients' ability to express their symptoms, moods, and cognitions honestly and willingly. Consequently, collecting diagnostic clues is a time-consuming procedure and involves a significant amount of clinical training, experience, and qualification to reach satisfactory results.

While bio-markers for both illnesses remain questionable, recent significant advances have been achieved in utilising affective computing and social signal processing as a screening tool, for example, using automatic speech and language analysis models to detect the changes in voice's characteristic that correlate with dementia or depression.

The speech analysis approach is a affordable, non-invasive and can be deployed remotely, which can be a candidate to be used as an objective detection tool. Furthermore, non-verbal para-linguistic base approach more likely to be useful when utilised in the primary health care in coordination with current tools, based on the fact that clinicians cannot capture the changes in the characteristics of speech signals due to the effect of both disorders.

In this regard, the work presented here explores a solution for an automatic detection and screening tool, that is based on the analysis of a person's voice.

1.2 Objectives

To address the limitations presented in the previous section, this thesis is concerned with the following research questions:

1. The feasibility of developing an automatic, low cost and simple system to aid the doctors in identifying early signs of dementia?
2. The feasibility of designing an affordable and simple system that assists the doctors in predicting the severity of dementia and monitor its progression?
3. The possibility of developing an objective tool that can be used to identify depression and estimate its severity?
4. The feasibility of developing an objective tool for screening bipolar disorder?
5. What sort of audio features and machine learning models that can be used in designing such systems? What type of features that have the potential to develop these tools with language-agnostic capability?

The research is focused on the exploration of signal processing and machine learning techniques for elicitation of clinically useful information from speech signals. Therefore the aims have been set to:

- Extract acoustic features that capture irregularities in a person's voice attributable to the presence of dementia or other related diseases such as FMD.
- Investigate different approaches that eliminate undesirable acoustic variability.
- Investigate how background noise can affect the quality of extracted features.
- Develop a system that can be used to monitor AD progression.
- Develop a new acoustic-based system that can analyse patients-doctors conversations to infer dementia related symptomology.

- Derive new acoustic features useful in detecting a depression state.
- Develop an objective language-agnostic depression screening tool.
- Develop an automatic screening tool for bipolar disorder.

1.3 Thesis contributions

1. **Simple and robust audio-based detecting of biomarkers for Alzheimer's disease (AD)**.

   Many dementia screening methods rely on relatively computationally heavy processing involving speech recognition and the use of natural language processing techniques to achieve some degree of speech understanding at the linguistic level [33–36]. This makes them unsuitable as low cost home-based solution and means they are expensive to port to new languages. The proposed alternative solution investigates audio-only processing to address this challenge. The method is solely based on acoustic features and therefore would only require simple readily available audio technology that can be adapted to suit patient requirement either in terms of being portable or/and wearable. This system was evaluated using speech recordings from the DementiaBank corpus for subjects performing the Cookie theft Picture description task. The complete system is presented in chapter 4.

2. **Detecting and predicting Alzheimer's disease severity in longitudinal acoustic data**.

   This method investigates the deteriorating speech signal of people suffering from AD. The aim is to firstly predict a common clinical examination score used for dementia (MMSE) using acoustic information extracted from people describing a picture a common dementia assessment task. Secondly, the aim was to develop a diagnostic tool able to distinguish people with AD from people with Mild Cognitive Impairment (MCI) and healthy control (HC). This is done on a longitudinal dataset (DementiaBank) allowing us to study the natural deterioration happening across a total of three visits. A total of 811 acoustic features were extracted and used to build two machine learning models: a regression model capable of predicting MMSE scores for each visit with an average mean absolute error of 3.1, and a classification model with an average cross-visit accuracy ranging between 89.2%

and 92.4% when doing pairwise classification between the AD, MCI and HC classes. This work is presented in chapter (5).

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

   Neurodegenerative diseases causing dementia are known to affect a person's speech and language. Part of the expert assessment in memory clinics therefore routinely focuses on detecting such features. The current outpatient procedures examining patients' verbal and interactional abilities mainly focus on verbal recall, word fluency, and comprehension. By capturing neurodegeneration-associated characteristics in a person's voice, the incorporation of novel methods based on the automatic analysis of speech signals may provide more information about a person's ability to interact which could contribute to the diagnostic process. This project demonstrated that purely acoustic features, extracted from recordings of patients' answers to a neurologist's questions in a specialist memory clinic can support the initial distinction between patients presenting with cognitive concerns attributable to progressive neurodegenerative disorders (ND) or Functional Memory Disorder (FMD, i.e., subjective memory concerns unassociated with objective cognitive deficits or a risk of progression). The discriminative power of purely acoustic approaches could be integrated into diagnostic pathways for patients presenting with memory concerns and are computationally less demanding than methods focusing on linguistic elements of speech and language that require automatic speech recognition and understanding. This work is presented in Chapter 6.

4. **Language independent characteristics aid in depression evaluation**

   This work presents a new approach to the assessment and evaluation of depression, utilising temporal speech characteristics. The frequency and duration of voice segments and pauses were intensively used as identification markers in a variety of mental disorders; however, the newly derived features provide extra depth and capture a speech behaviour useful in discriminating depressive voices. Fusing these features with state of the art acoustic measures improve the model performance. The proposed approach evaluated using three depression data-sets and two spoken languages (German and English). Outperforming the baseline challenge results and comparable to more complex modalities that combined both speech and video features. The complete system is presented in chapter 7.

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

   This work investigates the use of the newly extracted features that were developed in chapter 7, and evaluate their usefulness in estimating the severity of bipolar disorder. The proposed system was tested using AVEC-2018 dataset, which is the only publicly available dataset for patients suffering from bipolar conditions. The proposed system had shown effectiveness of predicting the three states of bipolar i.e. remission, hypo-mania and mania with an unweighted average recall of 66% and 53.7% for development and test partitions respectively, theses results were higher than the baseline audio and video modalities. This work showed the usefulness of the new extracted features in predicting bipolar states, also the proposed system is less complex compared to other modalities that required an ensemble complex approach for this type of tasks. Finally, the acoustic features used to develop this system, have the potential to be language independent thus the proposed system can be used in different languages. This system is illustrated in chapter 8.

1.4 Thesis outline

The rest of the thesis is organised as follows:

Chapter 2 Overview of both dementia and depression disorders, symptoms, their influence on speech characteristics and the current diagnostic practices.

Chapter 3 Review of the literature for studies that used automatic approaches to extracting clinically useful information from aperson's speech and language , part one concerning the recent studies that aimed to develop automatic screening/diagnostic tools for dementia diseases and other related disorders. Whereas part two reviews recent studies focused on developing objective tools for the detection of depression and/or estimate its severity.

Chapter 4 Describe the proposed acoustic only system that classify AD patients from healthy subjects. In this chapter, details for each unit of the proposed system were described with performance comparison to other models from the literature.

Chapter 5 Present acoustic based system to predict and detect the severity of Alzheimer disease in longitudinal dataset. The details for this system were presented along side the dataset that used to report the results. Also, the proposed system compared to other modalities from the literature.

Chapter 6 Describe the new proposed approach which evaluates people conversations with doctors at the memory clinic, as this system designed to capture dementia-related symptom through the analysis of the patient's voice. Finally, this model compared with a more sophisticated approach that used a combination of three types of features to produce a similar performance.

Chapter 7 Present the proposed depression screening system. The system evaluated using German and English datasets. Also, this chapter introduces the newly developed speech activity features. Finally, the system performance compared to other studies from the literature.

Chapter 8 Introduce an automatic system for screening Bipolar disorder. This system built using a combined features from the newly developed set and from the state of the art acoustic features. The proposed method evaluated using the AVEC-2018 Bipolar dataset spoken in Turkish. Finally the system performance also compared to other studies from the literature.

Chapter 9 Includes the summary with conclusions and the further direction.

Chapter 2

Dementia and depression, symptoms and current diagnostic practices

2.1 Dementia

Dementia is a brain disease result from the damages affects the brain neurons' synapses; it's a progressive and irreversible disease, characterised by losing the cognitive functionality. People with dementia and over the time suffer from loss of memory, a decline in the understanding, problem-solving, poor judgment, weak communication skills [37], unexpected changes in moods, apathy, depression and limited motor control. At the severe stage, the patient loses the ability to speak and can't perform the simple life dependent functions, for instance, swallowing and eventually causes death [38].

Dementia normally targets elder people at age (60-65) and over [39]. There are many diseases that cause dementia, for example, Vascular Dementia (VD), Dementia with Lewy bodies (DLB), Frontal Lobe Dementia (FLD)and Parkinson Disease; however, the most common cause of dementia is Alzheimer disease (AD) which represents 60-80 % of dementias [7, 10].

AD caused by two tiny deposits known as Tangles & Plaques accumulated in the brain neuron's synapsis result in destroying the connections between the neuron cells, which contain the memories, sensation and the motor ability, thus the most common symptom of AD are progressive memory impairment which affects the visuospatial and executive functions as well as the individual's behaviour [40, 41].

The second commonest form of dementia, in the UK is VD, estimated at 20% of all dementia. VD is general term describes several conditions result from damages to brain's blood circulation. The damages may cause by small blood clots obstructing oxygen supply to brain cells, blocked arteries, and exploding of blood vessels in the brain. VD symptoms depends on which part of the brain is affected, and that may include: communications, language, reading, and writing. Although issues with memory may not be instantaneously presented, depends on the damaged part, it may appear afterward [42].

DLB is another progressive condition of dementia accounting at least (10-15)% it influenced movement and motor control. DLB caused by clusters of an abnormal protein called (Lewy bodies) accumulated in different parts of the brain. As a result, DLB has a variety of symptoms including sleep disturbances, hallucination, susceptible to fall, problems with swallowing, and tremors. However, memory is frequently less affected compared to other forms of dementia [43]. FLD is a disease result from loss of the nerve cells and pathways in both parts of the brain known as frontal and temporal lobes. The damage to the brain is associated with clustering of abnormal protein prevent communications between brain cells. FLD subtypes are; behavioural variant frontotemporal dementia (bvFTD), and Primary Progressive Aphasia which consists of Semantic Dementia (SD) and Progressive non-fluent Aphasia (nfPPA). FLD symptoms may include: changes in personality and behaviour, empathy, problems with decision making and concentration, identifying people or objects, apathy, speech and language problems [44]

2.1.1 Dementia progression and effect on language

It is believed that dementia may start more than a decade before it diagnosed; however, it is challenging to recognise the subtle changes were effects in the brain, in the cerebrospinal fluid (CSF), and the blood may begin; thus, the absence of the distinctive symptoms prevent the diagnosis at this crustal stage [45].

There are three stages of dementia which are clinically identified, early, middle, and late stages. The early-stage is known as "mild dementia" where patients' memory and cognitive capacities slightly reduced. Also, a subtle language deficit can be observed, such as difficulties in naming or finding the right word, troubles following a flow of conversation. Other mild declines could occur and may interfere with activities of daily living, including problems with organisations, planning, and visuospatial abilities. Patients will be aware of these changes as they developed, and as a result, depression, and anxiety symptoms also expected [46, 47]. These features can be subtle and may not be easily detectable. Specialist assessment is therefore valuable.

The next stage is the moderate stage; as dementia progresses, more symptoms may arise. Mood changes will be noticeable, and patients might suffer from fear, confusion, and increase memory loss. Communication and language become more challenging, and the patient may lose track of his/her thoughts, understanding what others said, and naming difficulties further progressed [47]. In the late (severe) stage of dementia, patients' memory will significantly deteriorate, their recent memories may completely forgotten, and often they unable to identify or confuse with family members. Patients also experience difficulties in maintaining concentration for a long time, increased disorientation (they will be confused with the place where they are, and with limited knowledge of the time). Communication at this stage, greatly affected, patients may not understand what has been said to them and often reply with incomprehensible words. Patients will use other methods to show their feelings and communicate their needs, for instance, behaviour, gestures, sounds and facial expressions [48].

Dementia is the most challenging medical condition in the 21^{st} century as there is still no cure and no treatment that delay or prevent its progression. However, several medications exist that help with dementia symptoms. For example, Donepezil and Rivastigmine inhibitors useful in term of enhancing the connections between the brain cells. These inhibitors boost the Acetylcholine level in the brain, which is vital neurotransmitter tends to diminish in demented subjected [49]. Memantine is another medication prescribed for those who in the moderate and severe AD stages. It is believed that Memantine can improve the dysfunctioned of glutamatergic neurotransmission (an excitatory neurotransmitter in the human nervous system) [50]. Other conditions can affect dementia symptoms, and that includes depression, high blood pressure and stroke.

In severe dementia stage, several behaviours were expected, such as aggression, illusions, anxiety, and loss of interest. Other treatments exist to help with these conditions, for instance, the cognitive stimulation therapies works at the early stage of the dementia, whereas the cognitive rehabilitation could help patients in mild to moderate dementia [51].

Various aspects of speech and language skills have been proven to be affected during the course of dementia. Subtle changes in the acoustic characteristics of the speech might occur in the early stage of dementia; these include, voice pitch, formants, shimmer, jitter, HNR [52] and spectral properties [34]. Speaking behaviour is shown to be affected as well, for instance, low speech rate, increase the number and duration of pauses and hesitations [53, 54]. Speaking behaviour is shown to be affected as well, for instance, low speech rate, increase the number and duration of pauses and hesitations [53, 54]. These effects might reflect the difficulties in finding the appropriate word (lexical deficits) or due to an increase in cognitive load (were the cognition already declined) or due to complications in planning for the next sentence. The linguistic characteristics reportedly to be affected as well [55].

2.1.2 Other conditions lead to memory defects

There is evidence showing that not all memory complaints caused by dementia; there are several conditions were patients share similar symptoms of memory problems to that caused by dementia [29]. These conditions have better chances to be cured comparing to those caused by neurodegenerative diseases (ND). However, identifying the actual cause is a challenging task due to the overlapping symptoms and the lack of accurate predictive biomarkers suitable for routine screening or stratification.

The mild cognitive impairment (MCI) frequently referred to as an early course of dementia; however, less than 15% of MCI patients develop to AD condition [45]. Whereas the majority of MCI conditions due to several factors, for example, drug use or issues caused by depression, all of which can be controlled; thus, symptoms can be improved [29].

In contrast to MCI, Subjective Cognitive Decline (SCD) is clinically referred to as an early form of dementia, although most of the patients referred to a memory clinic have subjective memory complaints and not SCD [29]. The subjective complaints may be increased due to self-awareness and education about the risk factors and symptoms linked to all forms of dementia diseases. Further, there are several psychological or pathological issues may attribute to these non-progressive subjective complaints [29].

Probably the most common cause of memory problems is known as Functional Memory Disorder (FMD). FMD is a clinical condition where patients have memory problems are not caused by an organic defect rather caused by a stressful event or due to psychological issues. FDM differ from SCD, as their complaints are pragmatic and not established on subjective bases

[56]. Furthermore, depression and similar to FMD can also lead to a non-progressive cognitive disorder. Although FMD and depression are associated with each other, one condition can cause another; however, a few percentages of FMD are depressed.

Finally, there is a group of patients that suffer from depression but do not have dementia, and these patients suffer from what is known as Pseudo-Dementia (DPD). DPD is condition were patients and especially the elder ones showing signs correlated with dementia, for example, lack of interest, attention, and frequently they answer with I do not know [57]. Furthermore, insomnia, lack of understanding, fatigue, sadness also frequent in DPD.

2.1.3 Dementia diagnosis

The current procedure for dementia diagnosis can be lengthy and complicated. The process includes a variety of steps that required examinations and specific tests. First, a standard checkup will take place to investigate if there are co-morbid diseases that can cause similar symptoms, for example, congestive heart failure. Next, a series of cognitive tests will be performed, in which memory, orientation, attention, language and executive function will be evaluated. At this point, further neuralological test may be necessary such as gate and/or cranial nerve examinations [58].

In addition to the procedure mentioned above, researchers have found several biological markers can assess the diagnosis process, and these include Cerebro Spinal Fluid (CSF), Computed Tomography (CT) scans, blood tests, Electroencephalogram (EEG), Positron Emission Tomography (PET), Single Photon Emission Computed Tomography (SPECT), Magnetic Resonance Imaging (MRI) etc. These tests can be cumbersome and costly as PET and also exposes patients to a high dose of radiation [59].

Furthermore, the neurologists take into consideration diverse of behavioural changes which might be observed during the diagnosis process such as walking, sleeping patterns and communication skills. For this purpose, a variety of tests exists to help to examine the individual's speech and language capabilities. When people visit a memory clinic, the assessment typically begins with a conversation with a specialist during which patients are asked a series of questions about their memory problems (known as a history-taking phase) then several screening procedures might follow. Most of the time this is done using pen and paper to mark each part of the test [60].

2.1.3.1 Cognitive tools

The existing tools for the cognitive test were designed to measure the patient's responses in a range of cognitive domains including place and time orientation, memory, language, visuospatial abilities, verbal fluency etc. Depends on which test has been used, the examiner and after completing the scores for each part, will compare the patient's final score with the test cut-off to estimate the patient's status (demented or no, and which stage he or she probably in). Below sections will briefly describe the most common tools that currently used. However, these tools have limitations, for example, how sensitive (true positive rate) or specific (true negative rate) they can produce, and other factors might influence the decision such as patient-level of education.

A. Minimal Mental Status Examination

The Mini-Mental State Examination (MMSE) is widely employed in clinical and research settings to assess cognitive impairment with (5-10) minutes administration time, developed by Folstein *et al.* [61]. It is also used to screen for dementia and to estimate the severity and progression of cognitive impairment. The tool can be deployed to monitor the cognitive changes longitudinally for an individual. MMSE measures the severity of cognitive impairment in five categories: attention, registration, language, orientation, and memory. The MMSE test consists of 11 questions with a top score of 30 points (a score over 25 considered a healthy cognitive state). The score below 10 reflecting a severe cognitive impairment, while a score between (10-20) indicate a moderate impairment and a score between (20-25) might considered mild cognitive impairment. Although MMSE test is widely used as a standard clinical screening procedure, it has several disadvantages; it is sensitive against age and years of education. Anthony *et al.* [62] suggest to use extra screening tools when testing patients who are more than 60 years old and have less than an 8-grade education. Also, it lacks sensitivity in the events of the progressive changes that occur to AD patients. Furthermore MMSE criticised for the low success rate to differentiate mild AD from healthy patients.

B. Montreal Cognitive Assessment

The Montreal Cognitive Assessment (MoCA) is another tool extensively used for evaluating cognitive impairment [63]. It was founded in 1996 by Ziad Nasreddine in Montreal, Quebec and required 10 minutes administration time. The top score is 30 similar to MMSE, but the cut off for normal state is 26. MoCA examines a variety of cognitive areas, including memory recall, visuospatial capabilities, language assessment, and orientation. MoCA have a higher success rate of detecting MCI compared to MMSE; however, its specificity is inadequate. Although it is available in 46 different languages, yet, it has several cut-offs that suggested to adopt the language and cultural differences [64].

C. Addenbrooke Cognitive Examination

The Addenbrooke's Cognitive Examination (ACE) [65] mainly implemented to increase the screening performance of the MMSE; also, it tries to address the neuropsychological omissions. ACE targets five cognitive domains: memory, visuospatial abilities, language, verbal fluency, and orientation/attention. It has a top score of 100, with suggested cut-off scores of 83 and 88, the higher scores mean better cognitive state. ACE and its following versions (Addenbrooke's Cognitive Examination-Revised, ACE-R[66] and Addenbrooke's Cognitive Examination III, ACE-III) [67] are neuropsychological tests utilised to discriminate cognitive decline in conditions such as dementia. Although ACE may have better sensitivity compared to MMSE, however, it demands extra time and an expert such as neurologist to administer the test.

D. Boston Naming Test

The Boston Naming Test (BNT) created in 1983 by Edith Kaplan, Harold Goodglass and Sandra Weintraub [2]. It is a neuropsychological test, which examine the confrontational word retrieval in individuals with various conditions including dementia diseases, aphasia, and other language defects cased by stroke. The test consists of naming 60 pictures that sequentially displayed to the patients. The examiner will evaluate the patient's responses using special codes, and in the event of patient failure to name a picture, the examiner may give a clue to help the patient. In the end, all responses will be processed, and the final score will be produced. The BNT depends on the individual performance generated from the successive items, and some pictures may not capture a monotonic growth in psychometric difficulty, some are insufficient to discriminate individuals at different stages of naming ability; also, multiple items provide redundant psychometric information [68].

E. Wechsler Adult Intelligence Scale

The Wechsler Adult Intelligence Scale (WAIS) is one of the most used tests to measure intelligence quotient (IQ) and cognitive ability in adults. The most recent used version of the test is the (WAIS-IV) released in 2008 [69]. This version includes ten cores subtests and five supplemental subtests. The four main indexes of the test are: Working Memory Index (WMI), Processing Speed Index (PSI), Perceptual Reasoning Index (PRI), and Verbal Comprehension Index (VCI).

F. Wechsler Memory Scale

The Wechsler Memory Scale (WMS) is a neuropsychological test developed to evaluate various memory functions [70]. The current version is the 4^{th} edition (WMS-IV) created in 2009 [71], and which was planned to be used with the WAIS-IV. WMS-IV contains seven subtests: General Cognitive Screener, Symbol Span, Design Memory, Spatial Addition, Visual Reproduction (I and II), Logical Memory (I and II), and Verbal Paired Associates (I and II). The individual's performance is generated as five Index Scores: Auditory Memory, Visual Memory, Visual Working Memory, Immediate Memory, and Delayed Memory.

2.2 Depression

Depression is a psychiatric disorder characterised with a diversity of symptoms: including continues feeling of sadness, negativity, lose concentration, difficulty in thinking, sleep disturbances, increase/decrease in weight, and people with depression often had the feeling of hopelessness, dejection, and sometimes suicidal thoughts [72]. Diagnosing depression is not a straight forward problem due to the variety of clinical profiles associated with depression. Evidence had been shown that individuals could have the same diagnosis in spite of having different symptoms [73]. The current standard tools for diagnosis depression are in the forms of questionnaires, which at risk of providing disappointing results. The false positives and false negatives might occur due to several factors related to subjective bias. To increase the performance of the existing diagnostic methods, objective assessment tools based on physiological, behavioural, and biological markers urgently needed [74].

The objective markers could have a variety of possible benefits lead to enhance the current diagnostic systems, for example the concept of a new therapeutic apparatus that provide

Table 2.1 Depression common symptoms.

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

instant evaluation, or directly contact the health services in the scenario of extreme depressive episodes. Furthermore, simply and remotely monitoring of depression conditions will be highly appreciated whether by a primary or the secondary health services providers [75].

Based on the fact that clinicians based their assessment of depression using symptoms related to the changes affecting an individual's lifestyle for example the behavioural changes, feelings,etc. Conversations with clinicians expressed those. Therefore speech has the potential as a prominent objective diagnostic tool. Research into this direction requires understanding how the current diagnostic procedures works, what other sorts of physiological, biological, and behavioural markers used to aid the diagnostic system. Speech production mechanism influenced by many factors, mapping the aforementioned depression's markers that alter speech characteristics to depression markers is a key to achieve our goal.

2.2.1 Clinical depression

clinical depression or major depression disorder is cased by variety of factors that impair the functionality of the cortical, sub-cortical, and limbic systems. Although those factors may not yet characterised, they possible to be linked with genetic amenability and surroundings conditions such as stress and emotional shock [76–78]. According to Statistical Manual of Mental Disorders (DSM), depression is diagnosed when an individual suffers from either depressed mood (DM) or anhedonia (loss of interest or pleasure) associated with at least four or more symptoms listed in Table 2.1 and for at least two weeks period [79].

The DSM is extensively used manual in the diagnostic of mental disorders, and its current fifth version published by the American Psychiatric Association in 2013. DSM was implemented to provide standard criteria for discriminating disorders based on perceived symptoms. It is, however, been criticized for the increased homogeneous groups compared to the previous versions (more disorders listed and some disorders were divided into subgroups or subtypes). Additionally, it is more challenging to reach a diagnosis in DSM due to its way of defining the boundaries between the groups and subgroups and therefore making diagnosis dependent on subjective biases where a valid patient examination should not take place to reach a diagnosis [80].

2.2.2 Diagnosing depression

Depression diagnosis in primary care settings is a challenging task. Applying the DSM criteria of depression (Table 2.1), Østergaard *et al.* [73] estimated that there is a large number of different profiles of depression; this variety adds an extra level of complexity while attempting to relate the clinical profile of a depressed individual into an objective level [81]. Also, the diagnosis might even harder for the clinicians due to the increased rate of individuals seeking aid in primary care, the considerable amount of time it takes to obtain a diagnosis, physical signs concealing their underlying cause, increasing the risk of misclassification, and some depressed patients may not willingly be able to reveal emotional symptoms such as sadness or hopelessness [82].

Depression tools for screening patients exist, whether inform of clinical-led or self-evaluation questionnaires; they aim to measure the severity of symptoms through a scoring mechanism. This mechanism is sensitive to all underlying depression symptoms, make it vulnerable to subjective bias due to some symptoms; for example, low mood is not physically measurable. This bias risk required extra training in clinicians-led to achieve satisfactory results [83, 84]. Moreover, in order to reach the full potential of these evaluation systems, the patients have to be able to express their symptoms of feelings, moods, and cognitions willingly and honestly, which it might not be the case always due to the impaired motivations, and outlook from the first place.

2.2.2.1 Depression evaluation tools

The screening tools used in the depression speech databases which used in this study are the Hamilton Rating Scale for Depression (HAMD) [85], or self-evaluations such as the Beck Depression Index (BDI) [86] or the Quick Inventory of Depressive Symptomatology (QIDS) [87]. These tools represent the gold standard in rating the severity of depression symptoms. The resulting score indicates the level of depression's severity, and this score generated differently across these tools, different weighing methodologies and variety of symptoms they tend to cover. However, they share several common symptoms including; suicide thoughts, loss in interest/pleasure, increased feelings of guilt, changes in appetite, increased fatigue, changes in sleeping patterns, increased depressed mood and increased agitation.

A. Hamilton Rating Scale for Depression

Hamilton Rating Scale for Depression (HAMD) is considered to be the gold standard tool for assessing depression severity. First it was introduced in 1960 [85], and then was revised couple of times in 1966 [88], 1967 [89], 1969 [90], 1980 [91]. It is a multiple item questioner created to screen adult for depression, and measure the severity of their condition by investigating various aspects of the individual's day to day feelings of guilt, mood, suicide thoughts, anxiety, somatic symptoms, agitation or retardation, and insomnia. HAMD consists of 21 questions and required administration time between (20-30) minutes. Each question has 3-5 outcomes depend on the severity of the perceived response. The test has five evaluation levels; normal (0-7), mild (8-13), moderate (14-18), severe (19-22), and very severe (≥23). The HAMD has been frequently criticised for its inability to rank patients in terms of their severity status. [92, 93].

B. Beck Depression Index

The Beck Depression Index (BDI) is one of the most extensively used self-assessment tools of depression [86]. It encompassed 21 multiple items, that emphasis on the somatic symptom, and essential cognitive noticed in depression alongside the negatives on self-evaluations such as self-criticisms and self dislikes. Each item marked as 0, 1, 2, or 3 estimated based on how the patient interprets the severity of a specific symptom over the last week. The top score is 63, and it has four scales of depression: severe (≥30), moderate (19-29), mild (10-18), and minimal (0-9). The tool considered to be convenient to use (short self-assessment, no need for clinicians) and reliable [94]; however, the patient's reading ability and learning effects influenced its reliability [95].

C. Quick Inventory of Depressive symptomatology

The Quick Inventory of Depressive Symptomatology (QIDS) is a brief self-reported test for measuring the severity of depression, and it takes 5-10 minutes to complete [87]. It was developed based on a previous lengthy test (takes 20-30 minutes administration time) known as the 30-question Inventory of Depressive Symptomatology (IDS) [96]. There are two existing versions of QIDS, the self-report (QIDS-SR_{16}) and clinician-rated (QIDS-C_{16}). Both versions have scores range from 0 to 27, the final score is accumulated based on the patients' responses to the 16 different items. Those items summarised into nine questions, and each question can be scored between 0 and 3. The nine evaluation categories are: interest, sleep disturbance (initial, middle, and late insomnia or hypersomnia), sad mood, psychomotor, agitation/retardation, energy/fatigue, self-criticism, suicidal thoughts, concentration, and changes in appetite/weight. The test has five evaluation levels: very sever ($\geq$21), sever (16-20), moderate (11-15), mild (6-10), and normal (0-5). QIDS considered to be a reliable tool for depression assessment, however, it more frequently placed patients into the highest level of severity compared to other measures, and that requires medications and treatments which may not be needed at the first time [97]

D. Montgomery – Åsberg Depression Rating Scale

The Montgomery – Åsberg Depression Rating Scale (MARSD) was created in 1979 as a supplement to HAMD scale [98]. It is a ten questions tool used by psychiatrists to estimate the severity of depression in individuals suffers from depressive episodes associated with mood disorder [99]. MARSD expected to have higher sensitivity than HAMD in the events of the changes that occur due to the antidepressants and other forms of treatment. The score can be any value between 0 to 60, and a higher value means more severe depression. The ten questions measure various symptoms which are: reduced sleep, inner tension, concentration difficulties, pessimistic ideation, reported sadness, apparent sadness, suicidal thoughts, inability to feel, lassitude, and decreased appetite. Each question can have score ranges from 0 to 6, and the total score indicates one out of four levels of depression the normal (0-6), mild 7 to 19, moderate (20-34) and severe depression >34. The MARSD takes 20-30 minutes to complete.

E. Zung Self-Rating Depression Scale

The Zung Self-Rating Depression Scale (SDS) implemented in 1965 to examine the severity of depression for patients suffering from depressive disorder [100]. It has 20 items and takes 5 to 10 minutes of administration time. Each question can be scored from 1 up to 4, and final raw scores range from 20 to 80, which need to be converted into depression severity index (multiplied by 1.25 and divided by 100). The index has four levels; normal (20-44), mild depression (45-59), moderately depressed (60-69), and ≥70 for severely depressed. While SDS is proven to be a reliable test, it was reported that the test might have an unsatisfactory correlation with age. The non-depressed individuals whose age is less than 19 years, and older adults above 65 years of age their scores lean to fall in the depression category [101].

F. Zung Self-Rating Anxiety Scale

The Zung Self-Rating Anxiety Scale (SAS) is a self-evaluation test used to estimate anxiety levels [102]. The test designed to assess four aspects of an individual symptom's: motor, central nervous system, automatic and cognitive. The responses are evaluated in one or two weeks before the test. The answers are scored between 0 to 4 based on the following choices: "a little of the time", "some of the time", "good part of the time", "most of the time". The final score is a sum for all replies from the twenty questions, and this raw value scaled into anxiety index score, which ranges from 20 to 80. The normal range (20-44), mild to moderate (45-59), the severe anxiety (60-74), and extreme severity is ≥75. Although the SAS is a brief test can be completed in 5 minutes, it criticised for not showing adequate discrimination ability between depression and anxiety [103].

G. Patient Health Questionnaire-9

The Patient Health Questionnaire-9 (PHQ-9) is test used to measure the severity of depression [104] with top score of 27, score over 20 indicate sever depression, 15 to 19 score is a moderately sever depression, a score between 10 to 14 is mild depression, 5 to 9 score is moderate, while score up to 4 is the normal or no depression. The test is self administered with 9 questions, each question score ranges from 0 to 3. PHQ-9 target emotion, feeling, interest and satisfaction in individuals.

H. Generalised Anxiety Assessment-7

Generalized Anxiety Disorder 7 (GAD-7) is a self-reported questionnaire for examining and measuring the severity of generalized anxiety disorder (GAD). The GAD-7 test top score is 21, and each one the seven questions can be scored between 0 and 3. A Score of more than 15 means severe anxiety, score range from 6 up to 14 is moderate-severe anxiety while mild anxiety score between 0 up to 5 [105].

2.2.3 Depression objective markers

Using behavioural signals or measurable biomarkers in the diagnosis process is not yet entirely adopted in psychiatry utilities. However, research in these directions becomes attractive among the researchers. It is still a challenging task to find a specific objective marker given the vast diversity of symptoms associated with depression [73]. Combining potential speech signal signs with other behavioural and physiological markers could be one way to achieve desirable results. Therefore, it is important to briefly review the recent research into the related behavioural, physiological and biological markers.

2.2.3.1 Physiological and biological markers

Defining depression-related psychological and biological markers is the core of research towards an objective assessment tool. It has been reported that several biological markers associated with depression, for example, genetic abnormalities [30], low serotonin levels [31] and neurotransmitter dysfunction [32]. However, to date, no particular biomarker has been identified. This limitation might be due to the variety of symptoms, absence of strong relations between genes and depression pathologies [106]. The low level of molecular serotonin is considered to be the best biomarker for depression indication [106]. However, several reports showed that healthy subjects from a family with mental illnesses [107], or subjects suffer from aggressive behaviour [108], or suicidal subjects [109], might as well suffer from low serotonin. Therefore, low serotonin is not entirely a depression specific marker.

Depression could be investigated through neuroimaging, as it found that several interactions provide useful information for this purpose: the interactions between the cortical and limbic system linked with onset depression [110, 111], interactions between the gene responsible for brain-derived neurotropic factor (BDNF) and recurrent of the stressful event have been associated with risks of anxiety and depression disorders [30], the reduced activity of neurotransmitter gamma-amino butyric acid (GABA) correlated with the risk of having depression [112], additionally, GABA transmission linked with progressed stress activity [113].

The motor co-ordination controlled by the basal ganglia (area in the brain) is connected to the prefrontal cortex and limbic system, which means basal ganglia affected through the course of depression [114]. Depression also mapped to the size of hippocampal [115]. Frodl *et al.* [116] reported that 60 patients suffering from depression have small hippocampal compared to matched healthy subjects. The authors also liked between the BDNF gene and the small volume of hippocampal; this could mean the gene abnormalities is a marker for developing depression.

Additional biomarkers linked to depression are; cytokines which is a protein molecule [117], insulin and serum molecule [118], and level of salivary cortisol, a steroid hormones [119]. while Phsiological markers encompass: dysregulation of cardiovascular [120], galvanic skin reactions [121] , saccadic eye motions [122], and sleep distrubances [123].

2.2.3.2 Behavioural markers

Despite that, the current diagnosis tools of depression are not focusing on adopting non-verbal communication and other behavioural markers, recent studies showed that these markers could be useful in discrimination depression symptoms, for example using eye gaze and eye movment [124], facial landmark tracking [125, 126], and hand gesture or movement [127].

2.2.4 Depression and speech

Emil Kraepalin [128] the most contributor to modern psychiatry, characterised depressed patient speak with longer pauses and hesitations, slowly and with low volume, and sometimes whispering. Depression influence speech, and this effect clinically often described as follows: reduced in speech rate and verbal activity, short utterances, and long silent pauses [74, 129].

Speech is the most available source of information and what is more that makes it an attractive candidate as an assessment tool is that; it is a noninvasive and non-intrusive approach, it can be remotely deployed and cheaply measured. The human brain controls speech production, and it is a sophisticated process; thus, it could be affected by subtle cognitive and psychological changes. Therefore the potential changes might be reflected in altering the speech acoustics's characteristics [130]. Under depression, both cognitive and psychological changes occur, and it expected to affect the controlling mechanism behind the speech production process. This alteration to the acoustics presumably measurable and possible to consider as an objective assessment approach.

2.3 Advantages of an automatic screening tool

Both dementia and depression diagnosis is not a straightforward process. Dementia symptoms are overlapping with a variety of conditions including (FMD, SCD), normal ageing and even depression, and all of that can produce either misdiagnosis (false positive) or under-diagnosis (false negative). Similarly, identifying depression states can be influenced by subjective bias and/or honesty and accuracy of patients expressing their symptoms. Besides, the tools that can be used to recognise individuals more-likely of developing dementia are invasive, expensive, and some tests expose people to radiation. Whereas other cognitive tests are non-invasive and cheap, yet their sensitivity and specificity somewhat sub-optimal and performed using papers and pen, also they produce learning effects which limit the number of possible administrations.

As a result, developing a cheap, non-invasive and non-intrusive, automatic, and objective tool that can be used frequently without learning effects, remotely applicable, reliable, and easily administrated is on high demand by healthcare providers. This tool can bring tranquillity for those at low risk of developing dementia, and at the same time to speed up the process of providing the right medications to those who most likely demented.

2.4 Summary

Dementia and depression are the two most common causes of mental disorders affecting human-life. Dementia is a brain disease result from the damages affects the brain neurons' synapses, it's a progressive and irreversible disease, characterised by losing the cognitive functionality. The most common cause of dementia is AD which represents 60-80 % of dementias. There are three distinctive stages of dementia: early, middle, and late stages. Dementia symptoms includes loss of memory, decline in cognitive abilities and mood changes and these become more disturbances as the disease progression. Depression symptoms on the other hand, includes aggressive behaviour and suicidal thoughts.

Various aspects of speech and language features are affected during the course of dementia and under depression condition. Subtle changes in the acoustic characteristics of the speech might occur in the early stage of dementia; these include, voice pitch, formants, shimmer, jitter, HNR. Speaking behaviour is shown to be affected as well, for instance, low speech rate, an increase in number and duration of pauses and hesitations these also characterised in depressive voices. These effects might reflect the difficulties in finding the appropriate word (lexical deficits). The linguistic characteristics reportedly to be affected.

There are other diseases or conditions, such as FMD can cause similar behaviours as in dementia, however, these conditions can be cured. Therefore it is highly desirable to accurately discriminate FMD conditions from other ND diseases. It is believed that dementia may start more than a decade before it diagnosed; however, it is challenging to recognise the subtle changes due to the lack of accurate biomarkers. The current dementia diagnostic tools includes (CSF, CT and PET), sleep and gate examinations in addition to cognitive screening tools such as MMSE and MoCA. These tools either invasive, costly, or have lack of frequent usage. Whereas, the screening tools for depression includes PHQ, BDI and HAMD, these tools heavily rely on the patient's honesty to express their symptoms and the examiner ability to interpret the observed symptoms, so subjective bias may influence the evaluation decision. Thus, it is highly desirable to find simple, reliable and cheap approach can be used as diagnostic tool.

The study investigate the use of acoustic only approach to design simple and cheap tools that can help the doctors in detection/screening dementia, depression and bipolar.

Chapter 3

Automatic approaches to extracting clinically useful information from a person’s speech and language

3.1 Introduction

This chapter presents a literature review of some of the most extensively used methods for processing speech signals to extract clinically useful information. The search will focus on the automatic systems that recently developed to overcome the limitations of the tools and tests that currently use in diagnosing dementia and depression. More specifically, this work will explore; first, the types of features used, how these features extracted, and in which context and type of disease they were investigating. Secondly, to describe the machine learning algorithms used to build these automatic systems, and how have these models been evaluated, which metrics were used to report the results and what datasets have been used. Finally, to identify the gaps and limitations in these studies.

The rest of the chapter is organised as follow section 3.2 provides literature for the recent studies aimed for the detection of the early signs of dementia. Section 3.3 describes the latest studies that proposed automatic screening systems for depression. The last section is the chapter summary.

3.2 Dementia

Several studies have suggested that analysis of speech and language may offer valuable clues to detect behaviour bio-markers of dementia and Mild Cognitive Impairments (MCI) [5, 45, 131]. The Speech production starts in the left hemisphere of the brain [37]; and it is a complex process involving various cognitive domains, such as attention, planning, and memory, besides the language itself. As a result, a subtle decline in speech capabilities has been spotted for a long time (years) before dementia is diagnosed [132–137].

Automated speech and language analysis methods are potentially powerful tools, especially when using machine learning algorithms capabilities to evaluate the features extracted from the speech. Many studies have already applied machine learning approaches to identify text and/or speech samples from individuals with cognitive impairment [54, 138–144]. However, the significant obstruction in this domain of research is the shortage of high-quality, clinically-validated data needed to train such machine learning models.

Numerous studies have explored the relationship between cognitive decline and various aspects of speech and language. The following will review the findings regarding the use of different speech-based tasks and other cognitive tasks, and conversations to detect signs of dementia diseases and other cognitive disorders.

3.2.1 Dementia detection using linguistic features

Several studies investigated language based features, Bucks *et al.* [53] explored how AD deficits the speech and language compared to normal people. They collected speech samples from the interviews (continued for between 20 and 45 minutes) with 16 participants (8 controls, 8 with moderate and severe AD). Several lexical features were extracted from the manually transcribed audio recording. The feature set includes noun, pronoun, adjective and verb rates, Brunet's index, Honore's statistic, and the semantic cohesion, those features widely used in measuring the speech quality, flow and fluency. All the features were analysed using principal component and linear discriminant analysis. The binary classification result was 87.5% using the leave-one-out cross-validation method. The authors showed that people with AD suffer from word-finding, low verbal rate, and difficulty to construct longer phrases in contrast to the control groups. Thomas *et al.* [138] suggested lexical based approach to identify and rate AD. The analysis applied to speech recordings collected from 95 patients and their caregivers, all participated in a study supervised by the Atlantic Canada Alzheimer's disease Investigation of Expectations (ACADIE). The ACADIE investigated the effect of donepezil hydrochloride treatment consumed by the patients for 12 weeks. The follow-up assessments administered at 12, 24, 36, and 52 weeks. The patient's MMSE scores were showing mild, moderate, and severe AD [145]. The author used two interviews recorded with examiners for each patient at week 12. Thomas *et al.* used a manually transcribed text files for the recordings. The extracted feature set included the rate of adjective, noun, verb and pronoun, type-token ratio, Brunet index, Honors statistics, and clause like semantic unit features. Four classification scenarios were implemented. The maximum accuracies reached highest at 69.6%, 50.0%, 94.5%, and 75.3%, when doing (normal or mild AD) vs. (moderate or severe) AD, (normal vs. mild vs. moderate vs. severe AD), (normal vs. severe AD), and (normal vs. mild AD) respectively. Roark *et al.* [54] found that the complex linguistic measures and pause statistic features provide useful discriminative markers which can be used to differentiate between MCI and HC group members. The authors used SVM classifier to measure the area Under Receiver Operating (ROC) curves as a classification accuracy, achieving a maximum ROC measurement of 0.86 when selecting 32 (language and speech) features combined with neurologist test scores. Ahmed *et al.* [146] investigated the deterioration of language characteristics during the course of the disease for 15 subjects AD confirmed postmortem. The analysis applied to speech samples recorded while the patient was performing the Cookie Theft picture description task. The study reported subtle language defects observed during the preceding stages of AD (mild and moderate cognitive impairments), however, as the disease progression to AD stage, the linguistic measures including lexical content, semantic and syntactic complexity clearly altered through the consecutive clinical stages of the disease. Asgari *et al.* [143] conducted linguistic analysis to differentiate between two groups of subjects, first group represent cognitively intact subjects(n=27, age=78.9, and MMSE=28.7), and second group of patients having MCI diagnosis (n=14, age=83.4, MMSE=26.9). Both groups interviewed in standardized pre-selected topics, the conversations were recorded and transcribed for linguistics feature investigation. They used the Linguistic Inquiry and Word Count (LIWC) based features with a size of 68-word subcategories, for example, filler words, nonfluencies ('um,' 'er'), job-description words such as "Boss," "employee," etc. SVM and random forest classification algorithms trained under five-folds cross-validation method, and used to report the results. The highest classification accuracy was 76.2% achieved by non-linear SVM with radial basis function compared to 74.4% for the random forest classifier.

3.2.1.1 Linguistic approach applied to DementiaBank dataset

All previously mentioned studies used their own datasets and these are not publicly accessible, thus it is hard to compare the performance of this study with their modalities. However, many studies proposed systems for the same task and used the DementiaBank corpus [147] to report the results. The DementiaBank is the only publicly accessible dataset for dementia and relate diseases. Using language based modalities, Orimaye *et al.* [33] proposed a diagnostic method to identify people with AD using nine syntactic and eleven lexical features extracted from transcribed audio files from the DementiaBank dataset. They used a sample size of 242 files for both healthy older people and people with AD. They explored four different machine learning classification algorithms, achieving a 74% classification accuracy using a support vector machine (SVM) classifier with 10 folds cross-validation. The recent work by Orimaye *et al.* [148] used only 99 transcripts from the first visit by those with probable AD conditions, and matched it with 99 from the HC. The authors used the same feature set from the old work [33] and added N-gram (bigrams and trigrams) based features. They used the Sequential Multiple Optimisation (SMO) with RBF kernel model to test the accuracy. The authors used the Area Under ROC curve (AUC) to report the performance. The max AUC reported was 0.93 using the combined top 1000 features. Zhou *et al.* [149] utilised ASR Word Error Rate features for the detection of AD subjects, they experimented on the DementiaBank dataset which contains 167 AD participants provided 240 speech samples, and 233 additional recordings belong to 97 HC subjects. They only used lexical features generated from the ASR output and used to train the SVM classifier. The study struggled to provide a reliable accuracy due to the challenges of the noisy recordings, and the best WER obtained was 38.24%, this study shows that performance of such models strongly relays on the efficiency of the ASR. Hernandez *et al.* [36] used speech recordings from the DementiaBank dataset, with total of 517 samples divided into three groups (AD =257, HC = 217, and MCI =43). The authors used manually transcribed text files to extract a set of linguistic features. They include part of speech distribution variables such as frequency and ratio of nouns, verbs, conjunctions, vocabulary richness features, etc. Also, they extracted several statistic features applied to the Mel Frequency Cepstral Coefficients (MFCCs). With a total of 105 linguistic and acoustic features used as input to train SVM and Random Forest (RF) algorithms. The ten-fold cross-validation approach used to report the results. The correlation-based approach was adapted as a feature selection method. They selected only features that have a high correlation value with respect to the class. The SVM algorithm performed better than RF in a binary classification task ,i.e., identifying HC subjects from AD groups, with an accuracy of 79.0% compared to 75.00%. However, when classifying HC subjects from both (AD and MCI), RF outperformed the SVM and achieved an accuracy of 77.0% compared to 75.0%.

Further language approaches, Ammar *et al.* [150] proposed linguistic-based approach to diagnose AD from HC subjects. Using the manually transcribed text files the TalkBank CHAT (Codes for the Human Analysis of Transcripts) protocol [151], a total of 28 syntactic, semantic and pragmatic features extracted from the DementiaBank speech recordings. These features include the total number of nouns, verbs, adverbs, adjective, utterances, and the number of past and present participles, prepositions, conjunctions, pronouns, and error words, type token-ratio, etc. With the help of KNN-classifier based wrapper feature selection method, 11 most informative features were selected and used to train an SVM classifier. The 10-folds cross-validation procedure achieved precision at 0.79%. Klumpp *et al.* [152] used the same 499 transcriptions from the DementiaBank dataset, to construct an Artificial Neural Network (ANN) model to distinguish between ADs and HCs. They used the bag of words parameters and estimated 546 values that measured the occurrence of stemmed words following and using the Stanford CoreNLP Toolkit [153]. The fully connected feed-forward ANN designed to have one input layers followed by a hidden layer both having 546 nodes, while the output layer consists of two nodes representing the two classes. The authors feed the ANN using a batch of 20 samples and adopted 80% dropout of the values between the layers as a precaution to the over-fitting issue. All nodes configured to have Rectified Linear Unit (ReLU) activation function. ANN's best classification score was 84.4%. Mirheidari *et al.* [154] investigated word vector representations (WVR) features to classify between HC and AD. The model tested with samples obtained from the DementiaBank dataset. The authors compared two deep neural networks models that were built using WVR based features. The first model was built using the manual transcripts, and the second by using an ASR. The highest score achieved by the model that used the manual transcripts with a classification accuracy of 75.6% compared to 62.3% that used ASR. Fritsch *et al.* [155] proposed neural network language model to diagnose AD group from HC subjects. The model was built using statistical measures applied to a set of linguistic parameters. The authors based their analysis using statistical N-grams language models that means the probability for a sequence is defined by the product of the probabilities of the words that form the sequence. The model trained and tested with DementiaBank dataset, with a total of 499 samples comes from 168 ADs (having 255 recordings), and the rest(244) belong to the 98 HC participants. The authors used leave-one-speaker out cross-validation to test their model. The best classification result reported was 85.6% in distinguishing between the two groups. They also reported a high correlation between the MMSE scores and their features, both Pearson and Spearman's rank correlations were 0.656 and 0.771, respectively.

3.2.2 Dementia detection using acoustic features

The following studies proposed automatic systems based on acoustic features for this task, Lopez-de-Ipena *et al.* [156] investigated using features called Emotional Temperature derived from the speech along with acoustic features from 20 healthy subjects and 20 people who have dementia. This was done in an attempt to evaluate the importance of the emotions encapsulated in spontaneous speech, and they showed promising results when attempting to differentiate different stages of the disease. Later same group, Lopez-de-Ipena *et al.*[157, 158] examined a combination of linear and nonlinear acoustic features derived from a spontaneous speech in a multi-lingual dataset of 70 participants. The features were used to build a machine learning model designed to capture the irregularities affecting speech caused by AD. Melan *et al.* [52] used speech recordings from 30 AD patients (mean MMSE score = 18.07, mean age = 78.66 years, mean years of education = 6.27) and 33 HC subjects (mean MMSE=27.97, mean age = 74.06, mean years of education = 7.30). The recordings were collected while the participants were reading two Spanish sentences that appeared on the screen. The authors extracted a set of acoustic features using Praat software[159]. The features included total duration, phonation time, Noise to Harmonic Ratios(NHR), speech and articulation time, F0 statistics, pulses, periods, mean periods, and various measures for the frequency (jitter) and amplitude (shimmer) fluctuations. The statistical analysis applied to the feature sets shows a significant difference between AD and HC in five features; the percentage of voice breaks, the number of periods, and voice breaks and NHR. These important features used by the discriminant function and obtained 84.8% classification accuracy. However, the accuracy dropped to 83.3% when applying a cross-validation procedure.

Whereas several studies proposed modalities for the identification of MCI condition, König *et al.* [144] used recordings from three groups of subjects identified as AD (mean Mini-Mental State Examination (MMSE) score 19), healthy elderly controls (HC) (mean MMSE 29) and (MCI, mean MMSE 26). The participants were instructed to carry out four short vocal cognitive tasks. For each task, a number of acoustic features were extracted and used to train a support vector machine (SVM) classifier. Three different identification scenarios were tested: HC vs MCI, HC vs AD and MCI vs AD with reported classification accuracies of 79% ±5%, 87% ± 3% and 80% ±5% respectively. Toth *et al.* [160] proposed a speech-based method for the early diagnosis of AD. The approach was evaluated using German speech recordings from 84 participants performing three tasks: immediate and delayed recall, in addition to spontaneous speech. The mean MMSE scores for the MCI (n=48) participants was 26.9 while the HC (n=36) group had mean MMSE scores of 29.1. After watching two short films, the participants were instructed to talk about the contents of these movies, once immediately after the end of the first movie, and secondly, after a 1-minute delay or a distraction for the last one. The spontaneous speech task involved recordings of the participants describing their previous day. From participants' speech prompted by these tasks, Toth et al. extracted a set of acoustic features measuring the number and average length of silent pauses and filled pauses, as well as rates for both speech and articulation. These features were used to build three classifiers to differentiate between MCI and HC. The best accuracy was achieved with a Random Forest classifier with an F1 score of 78.8%. Satt *et al.* [161] conducted experiment on 89 participants diagnosed as (19 healthy adults, 43 with MCI, 27 with AD). The subjects were instructed to complete four vocal tasks describe a picture while looking at it, repeat a sentence and syllables, and look at a picture and then describe it from memory. Total of 25 vocal features were extracted for example the mean relative sentence duration, mean verbal reaction time, statistics of pauses and speech segments, the total number of the tokens uttered, mean number of errors per token average. With the use of single-tailed p- values, features that corresponding to the lowest value considered to be significant and therefore selected in estimating the classifier performance. The SVM classifier trained using 4-folds cross-validation method, and the reported results were in terms of Equal Error Rate EER. They used three configurations to test their method; control vs. MCI, control vs. AD and MCI vs AD, and the EER results were 17.00%, 15.5%, and 18.00% respectively.

3.2.3 Dementia detection using a combination of both linguistic and acoustic features

Using more demanding approaches, several studies had utilised both linguistic elements and acoustic features, Jarrold *et al.* [139] distinguished between different types of dementia by combining lexical and acoustic feature profiles extracted from spontaneous speech. The features were collected from 9 controls and 39 patients who had been diagnosed with different dementia sub-types: frontotemporal degeneration (mean MMSE 24), primary progressive non-fluent aphasia (mean MMSE 22), semantic dementia (mean MMSE 17) and AD (mean MMSE 18). The acoustic features included phoneme duration, speech rate, mean and standard deviation (STD) of the duration of consonants, vowels and pauses as well as the mean and STD for voice and voiceless duration segments. For the lexical features profile, they extracted the frequency occurrence of 14 different part of speech features (verbs, pronouns, nouns, function words, etc.). The feature-based profiles were derived from structured interviews and used as input to a machine learning algorithm. Classification accuracy of 88% was achieved by using a multi-layer perceptron as a binary classifier to differentiate between the AD and HC groups exclusively, while the classification accuracy dropped to 61% when all participants (with different types and severity of ND) were included. Weiner *et al.*[162] built a Linear Discriminant Analysis (LDA) classifier to perform a multi-class experiment on longitudinally collected speech samples. The dataset recordings were a collection of biographic interviews and cognitive tests of 74 participants administered by psychiatrists during the course of three separate visits. Follow-up cognitive tests demonstrated that some participants changed from the healthy cognition group into the Aging-Associated Cognitive Decline (AACD) and AD groups. A total of 98 speech samples were analysed (HC n=80, AACD n=13, and AD n= 5). A set of acoustic features were extracted using manually transcribed files and voice activity detector software. These features included the mean of silent segments, speech and silence durations, silence to speech ratio, silence count ratio, word and phoneme rates, and silence to word ratio. Using these variables, the LDA model achieved a classification accuracy of 85.7% between the three patient classes. Later the same authors Weiner *et al.*[163] added Automatic Speech Recognition (ASR) based features to their system. They also compared the performance of the automatic ASR based features to the manually transcribed conversations. Using both ASR types, they extracted several features such as perplexity feature, lexical richness, acoustic, ASR's word error rate, etc. The best result obtained was 0.623 using automatic ASR, reported as unweighted average recall (UAR). A year later, Weiner *et al.*[164] used Gaussian- acoustic-only model, and for the same data, the UAR dropped to 0.493, on the other hand, and when they used 241 interviews (collected from 218 participants), the model performance increased to 0.645 UAR when using only 12.5mins of the total conversations time. Sirts *et al.* [35] explored linguistics based features called propositional idea and semantic idea densities combined with other clustered features called LIWC [165] to distinguish AD patients from HC subjects. The authors used speech recordings from two datasets the DementiaBank and a dataset, recorded at Neuroscience Research Australia (NeuRA), contains interviews of autobiographical memory for both AD patients and HC subjects. The logistic regression classifier's results reported using F1-scores achieving 72.7, 78.4 for DementiaBank and NeurRA3 respectively, but when the clustered features trained using both datasets and added to the rest of features, the F1-scores increased respectively to 75.0 and 84.0. Gosztolya *et al.* [166] examined the fusion of two SVM models to perform binary, and multi-class classification to classify between three subgroups of subjects HC(n=25), MCI (n=25) and AD(n=25,mean MMSE scores 29, 27 and 23 respectively). The first model was built using a set of acoustic features which had been extracted using ASR software (articulation rates, utterance length, silent and filled pauses, a ratio of pause and speech). The second model was built using linguistic features extracted from manually annotated transcripts and included the number and rate of adjectives, nouns, verbs, pronouns, conjunctions, uncertain words, content words and function words. These features were extracted from recordings of Hungarian spontaneous speech. The accuracy of the fused model varied between 80% for both HC vs. MCI and MCI vs. AD and 86% for HC vs. AD, while the accuracy for the multi-class task, i.e. HC vs. MCI vs. AD was 81%.

3.2.3.1 Linguistic and acoustic models applied to DementiaBank dataset

The majority of studies used language based models and only few studies fused both linguistic and acoustic features, for example, Fraser *et al.* [34] studied the potential of using linguistic features to identify Alzheimer's disease. They used speech recordings along with their manually transcribed files derived from the DementiaBank data set. They chose 240 speech recordings belonging to a group of 167 people identified as probably or possibly having AD and 233 samples from 97 subjects with no memory complaint. In total, a set of 370 acoustic, lexical and semantic features were extracted, and they then applied two machine learning classification algorithms and obtained accuracy of 81% in distinguishing between HC subjects and AD patients using the top 25 ranked features. Later same group Fraser *et al.* [167] investigated multilingual linguistic-based features to classify between MCI and HC groups. The features set consist of cluster features (such as cosine distance between the centroid of all words that are members in that cluster), density and efficiency for information, idea, nouns, and verbs. The analysis applied to three datasets: Gothenburg [168], Karolinska, and the DementiaBank datasets. Total of 67 speech recordings utilized from Gothenburg (31 MCI and 36 HC) representing Swedish participants performing the Cookie Theft Picture description task. The Karolinska dataset [169] having only 96 HC subjects divided into two groups based on their age ranges (20-64) and (65-88). Both HC groups directed to produce a written description of what was happening in the Cookie Theft picture while having the picture in front of them. From DementiaBank two sets of samples designed, first set having (19 MCI and 19 HC) and the second set only have 78 HC subjects. The authors used a linear SVM algorithm to perform the classification task under the leave-one-out method. From the three datasets, different combinations were explored and used for training and testing. Best classification accuracies achieved was 0.63% and 0.72% when training the SVM using all samples from the all detests together but test only with the English and Swedish respectively.

Finally, predicting severity of AD through predicting the MMSE scores, Yancheva *et al.* [170] used a combination of acoustic and manually extracted linguistic features derived from the DementiaBank dataset, to predict MMSE scores. Total of 210 acoustic features included such as the mean and the mean of the mean for the first 42 MFCC coefficients and their skewness and kurtosis, pauses, word to pause ratio, the variance and average of the pitch and first three

Table 3.1 Summary of studies that proposed dementia detection systems.

Study	Dataset	Task	Evaluation metric	Accuracy

Buck et al.[53]	HC=8
	AD=8
		HC vs AD	Accuracy	87.5%

Thomas et al.[138]	95 subjects	(HC+ mild AD) vs (moderate + sever AD)	Accuracy	69%

Asgari et al. [143]	HC = 27
	MCI = 14
		HC vs MCI	Accuracy	76.2%

Melan et al. [52]	HC = 33
	AD = 30
		HC vs AD	Accuracy	84.4%

Koing et al. [144]	HC =15
	MCI = 23
	AD =26
		HC vs MCI
		HC vs AD
		MCI vs AD
			Accuracy	79%
				87%
				80%

Toth et al. [151]	HC =36
	MCI = 48
		HC vs MCI	F1_score	78.8%

Satt et al. [152]	HC =19
	MCI = 43
	AD =27
		HC vs MCI
		HC vs AD
		MCI vs AD
			Equal error rate	0.17
				0.15
				0.18

Jarold et al. [139]	HC = 9
	AD = 39
		HC vs AD	Accuracy	88%

Weiner et al. [153]	HC = 80
	AD = 18
		HC vs AD	Accuracy	85.7%

Gosztoyla et al. [157]	HC =19
	MCI = 43
	AD =27
		HC vs MCI
		HC vs AD
		MCI vs AD
		AD vs MCI vs HC
			Accuracy	80%
				86%
				80%
				81%

Orimary et al. [33]	DementiaBank	HC vs AD	Accuracy	74%
Ammar et al. [161]	DementiaBank	HC vs AD	Precision	0.79
Klumpp et al. [163]	DementiaBank	HC vs AD	Accuracy	84.4%
Mirheidari et al. [165]	DementiaBank	HC vs AD	Accuracy	75.6%
Fritsch et al. [166]	DementiaBank	HC vs AD	Accuracy	85.6%
Fraser et al. [34]	DementiaBank	HC vs AD	Accuracy	81%

formants, shimmer, jitter, linear predictive coefficients, autocorrelation, zero-crossing rate. The lexicosyntactic had a total of 182 features such as word type proportion, subordinate to coordinate phrases, type-to-token, mean length of utterance and ratio and Honoré's statistic. Finally, the semantic features included 85 variables. The authors used three feature selection approaches to find the most salient features, and these methods were the Spearman-rank correlation of features with MMSE scores, a t-test and the minimum redundancy-maximum-relevance. They achieved MAE of 3.8 when utilizing all samples with only 40 out of 477 features that were selected using the Spearman correlation method. The MAE improved to 2.9 for the scenario when only samples from patients who participated in more than three visits. In spite of the respectable results, their approach was based on the manually annotated texts, and it is unclear how the system behaves when adopting ASR in favour of a fully automated system.

The following chapters 4 and 5 will introduce the proposed systems and also compare the results with previously mentioned studies listed in Table 3.1 and specifically that used the Dementiabank.

3.3 Depression

Several studies also reported the use of speech-based systems for detecting mental illnesses. Depression disorder, for example, has been found to influence the prosodic, articulatory and acoustic features of a person's voice and language [171, 172]. The following are recent studies that proposed automated systems for the detection of depression disorder using speech, video and ensemble modalities.

The following studies used their own depression datasets, and therefore it is also difficult to compare between the proposed systems results and their approaches. However, there might be similar findings regarding depression investigation, which could be advantageous to healthcare providers when incorporating such systems to aid in depression diagnosis. These studies includes Moor *et al.* [173] applied statistical ANOVA test to build a classifier using quadratic discriminant analysis. This investigation demonstrated the suitability of glottal features for distinguishing between depressed subjects (n=15) and healthy controls(n=18). Although the study achieved maximum accuracies of 96% for female speakers and 91% for male speakers, running under leave-one-out cross-validation, the authors indicate that because of the small sample size (only 33 samples), these results might not generalise to a larger dataset. Low *et al.* [174] obtained classification accuracy varying between 70.3% to 77.8% for the binary male-gender, and between 67.1% to 74.7% for female-gender base tasks using speech samples from 139 adolescents performing a problem-solving task. The study showed that Gaussian mixture model (GMM) classification accuracy improved when combining features related to voice quality and Teager energy operator and spectral. Thus the authors, recommend adopting fusion at features level will likely improve the accuracy. Ooi *et al.* [175] used audio recordings from 30 adolescents, performed a discussion with their families regarding two topics event planning and problem-solving. Fifteen samples labeled as having no depression symptoms, while the rest have either major depressive disorder or another form of mood disorder. Total of 50 features was extracted and representing four types Glottal, prosodic, Teager energy operator (TEO), and spectral features. The authors used GMM classifier with weighted fusion accuracies, resulting from a different type of feature space, that reached a maximum value at 73%. Cummins *et al.* [171] compared the performance of two types of acoustic features MFCC and formants in classifying between depressed (n=23) and healthy (n=24) individuals. The analysis applied to speech recordings, collected while participants performing sentence reading task at Black Dog Institute. GMM classifier built with MFCC group of features achieved a higher classification accuracy of 77% compared to 74% when formant-based variables used. However, combining both types of features, GMM accuracy improved to 79%. Yang *et al.* [176] proposed model-based vocal prosodic features to estimate the severity of depression. The author used speech recordings from 49 patients participated in seven weeks of followup clinical trials for depression treatment. The average and coefficient of variation applied to the fundamental vocal frequency and switching pauses features which initially derived from the recordings. The Hierarchical linear modeling approach used to report the result, scoring 69% classification accuracy between low, mild, and moderate to severe groups. The study highlighted that switching pauses have more discriminating power as depression severity increases

Further study, Alghowinem *et al.*[177] used SVM classifier with leave-one-out cross-validation technique to differentiate between two groups of participants, HC group with 30 subjects and a depressed group having 30 patients. The participants were first interviewed, and during the conversation, a set of 8 open-topic questions used and later, participants instructed to read 20 sentences that have both positive and negative meaning. The author used speech recordings from the two tasks and extracted several features related to pitch, MFCC, energy, intensity, formants, and voice quality. The study reported that model constructed with features extracted using part of conversations provide better discrimination performance compared to a model built with features from the reading task. The SVM results reported as the weighted average recall, first model scores ranged from 60.0% to 78.3% compared to the second model scores ranged from 50.0% to 75.8%. Finally, France *et al.* [178] used speech recordings from three groups of participants labelled as health control (n=34), patient with high risks of suicide (n=43) and patients with major depression (n=42). The recordings from early sessions with a therapist, and after the therapy sessions were also included in the study. Various acoustic features derived from the recordings such as pitch and amplitude of modulation statistics, power spectral density, and formants and their bandwidth. The authors used a linear discriminant classifier to report the results. Four classification procedures were investigated control vs. major depression, control vs. suicidal, major depression vs. suicidal, and all classes achieved accuracies of 82.0%, 80.0%, 81.0%, and 75.0% respectively.

3.3.1 Depression evaluation using Audio/Video Emotion Challenge (AVEC) corpus.

The only accessible depression datasets are the Audio/video Emotion Challenge corpus (AVEC). These datasets will be described in chapter 7. In both datasets the AVEC-2013 and AVEC-2014, the challenge was to predict the severity of depression by predicting the Beck Depression Inventory (BDI) scores.

3.3.1.1 Depression evaluation using AVEC-2013 dataset.

Several studies found in the literature that used AVEC-2013, including a study by Williamson *et al.*[179] proposed GMM-based model for multivariate regression to predict the severity of the depression. The study used speech recordings from AVEC-2013 challenge, in which 340 video and audio recordings collected from 292 participants performing human-computer interaction tasks. The author used samples from the reading task only and used formants and MFCC based features (already provided with the dataset) in the experiment. With a multi-variant technique and principal component analysis, feature size reduced and fed to GMM regressor. The model obtained mean absolute error, and root mean square error of 5.75 and 7.4, respectively when using both feature types to predict the clinical depression scores, only evaluated with short time clipped from the reading passage from the development set only. Meng *et al.* [180] Utilised Motion History Histogram (MHH) to extract features from AVEC-2013 recordings. Both vocal and visual dynamics were used to capture changes in facial and vocal characteristics due to the effect of depression. The Partial Least Square (PLS) regression model used to evaluate the proposed approach in predicting the depression scores for both development and test sets. The authors investigated different modalities, including audio, video, and ensemble configuration. The audio model built using MHH applied to the 2268 audio features which supplied alongside with dataset. The video model constructed using MHH dynamics to capture temporal visual motion activities. HMM generate M gray-scale images for the video, which reflect the level of motion in the video, and then monitor the changes in the gray value occurred for each pixel for a duration of four consecutive frames. While the ensemble model used both audio and video modalities, Meng *et al.* used technique known as the linear opinion pool [181] to generate the decision rule. The PLS perform best when using the ensemble model achieved Mean Absolute error (MAE) and Root Mean Square Error (RMSE) of 6.94, 8.54 and 8.72,10.96 for development and test set respectively while the audio model only evaluated with development set and obtained MAE =9.78 and RMSE =11.54. The study obtained decent BDI scores prediction. However, it heavily relies on the provided set of features. Further, the decision fusion between the two modalities may generate bias as the authors did not report if the cut-off was the same for both development and testing sets. Cummins *et al.* [182] applied multi-models Gaussian Mixture with the universal background (GMM-UBM) combined with Support Vector Regression (SVR) algorithm to predict BDI scores for the AVEC-2013 corpus. The audio model contains MFCC coefficients and their delta and delta-delta derivatives, while the visual model constructed using features extracted from facial landmark tracking dynamics. The ensemble model has been built by concatenating audio and visual elements. Different kernels also investigated during the GMM-UBM models training and evaluations processes, such as the Kullback-Leibler (capture similarity between utterances) and the GMM-UBM Mean Interval Kernel (applied a covariance-based weighting scheme). Cummins *et al.* used cross-validation method for training and evaluation process. First, the model trained with the training set and then evaluated using five fold cross-validation, and the reported performance was the average of RMSE value cross all the folds. The best-reported RMSE scores were 10.44 and 10.17 for development and test sets, respectively. Both results outperformed the baseline challenge; however, the test configuration may generate a bias, the cross-validation folds mixed the training with test data. Further, the model tuning parameters were optimised during the cross-validation phase. a year later Cummins *et al.* [183] investigate the variability of acoustic volumes combined with Gaussian mixture model as indicator to speaker's level of depression and with same AVEC-2013 dataset, both development and test RMSE scores improved to 7.4 and 9.49 respectively. Kaya *et al.* [184] investigated the effect of applying three features selection techniques to improve the prediction of BDI depression scores. The authors used the pre-extracted audio-video features from the AVEC-2013 challenge to evaluate the proposed approach. They used Weka [185] tool to run the analysis. Among the three methods, the filter based-technique known as the "CFS" (Weka library that use the correlation between the features and the class to rank the variables) had performed better than the minimum redundant maximum relevant (mRMR) and the mutual information methods. The best-feature size selected was between 387 and 467 out of 2268. The SVR was used, and the best MAE and RMSE scores were 7.84 and 10.22, respectively. The authors didn't report the development results to see if the model performance is stable with a relatively large number of variables. Kachele *et al.* [186] Incorporated adaptive fusion method to improve the identification ability of the weak learner algorithm, which lead to in force the recognition of depression states. They extracted several audio-based features such as MFCC, rate of the glottal closure, skewness of glottal pulse and glottal harmonic method as well as video features such as appearance descriptors, and these features were derived from the AVEC-2013 dataset. Kachele *et al.* used Kalman filter technique to eliminate the uncertainty regarding the decisions created by the SVR algorithms, thus reduce the probability of false positives. A Multi-layer perceptron algorithm was constructed with three layers, the first one having 20 neurons with sigmoid output function, the second layer having five hidden layers(each having the same configuration as in the first one), while the last layer consists of 30 neurons. All modalities (audio and video) were evaluated individually and combined, the best results were with ensemble method achieved MAE and RMSE of (8.3,9.94) and (8.72,10.96) for the development and test sets respectively, while the audio scores (9.35,11.40) on the development and (10.35,14.12) on the test, and finally the video model obtained (7.03, 8.82) for the development and (8.97,10.82) for the test. Although the ensemble approach achieved better results, the audio model performs poorly in both test and that may due to the limited and the types of acoustics were employed in the construction process.

3.3.1.2 Depression evaluation using AVEC-2014 dataset

The following studies utilised the AVEC-2014, Simantiraki *et al.* [187] presented speech-based approach for the assessment of depression. The author utilised features called phase distortion deviation to capture irregularities affecting the phase component of the speech signal, which may be presented due to depression. The SVM classification algorithm used to test the model performance with AVEC-2014 dataset, and the results reported using the area under the receiver operating characteristic ROC curve. Features were extracted in addition to the voice pitch from both AVEC tasks reading and spontaneous speech. The two tasks evaluated individually and combined, and the best performance achieved was using both tasks and reached AUC 0.79 and 0.87 for female and male speakers, respectively. These results obtained for the development set only, and we don't know how the model will generalise to the test set of the same data. Further, it unclear how the proposed model will perform when a gender-independent scenario is evaluated. Mitra et al. [188] build two models namely multi-layer neural network and support vector models to predict the severity of depression (i.e. depression scores), the models were based on a combination of speech production, perception, acoustic phonetics, and prosody features extracted from AVEC-2014 dataset. The authors fused the two models to achieve MAE = 5.87 and RMSE = 7.37 in predicting depression scores for the development set. Pampouchidou *et al.* [189] combined facial and speech features extracted using open face and COVAREP tools respectively, to classify between depressed/non-depressed subjects, the method applied on the AVEC-2014 dataset achieving classification accuracy of 66% with weighted F1-score of 0.72, weighted precision of 0.94, and recall of 0.59 for gender based depression classification. The authors reported the results with development set only, and the best results achieve was based on the gender dependent. Morales *et al.*[190] compared the performance of two regression models that used to predict the Beck Depression Inventory-II (BDI-II) scores for AVEC-2014 corpus. The first regression model was build using speech based features. These features were statistical functions (percentage of frames loudness contour is above: minimum+25%, 50%, and 90% of the range, interquartile ranges, 1% percentile, standard deviation, kurtosis, arithmetic mean, root quadratic mean, minimum, maximum, skewness, quartiles, 99% percentile, percentile range 1%−99%) applied to features measuring the F_0, loudness contours, and voicing probability (which obtained using OpenSmile toolkit [191]), also this model included features related to the speech rate: average pause time, total duration, total speech time, average phone duration, total pause time, average syllable duration, and syllable rate. While the second model was built using text-based features consist of content features estimated using LIWC [192] dictionary. LIWC parameters incorporating: positive vs. negative emotion words, words referencing society,/friends/family, pronouns may capture inclusive language (we, us) vs exclusive style (they, them, you), and words referencing how the person is feeling (sleep, depressed, worried), another features also included related to part of speech tag n-grams and text-based speech-rate features using the Stanford Paresr toolkit [193]. The authors reported the performance in terms of mean absolute error (MAE) and root mean square error (RMSE), the first model had slightly better performance (MAE = 8.59, RMSE = 10.7) compared to the second model (MAE = 8.99, RMSE = 10.75). Morales *et al.*also investigated the effect of combining the two feature sets to build which improved the performance to MAE = 7.56, RMSE = 9.21 using Sequential Minimal Optimisation (SMO) regression model.

Further Sidorov *et al.* [194] investigated the audio, video, and a combination of both modalities to measure the level of depression severity in the AVEC-2014 development set. The performance evaluated using SVR algorithm. The authors used the provided audio features to build the audio model. While for video features they used eMax face analysis toolbox [195] to extract the common dynamic appearance descriptor known as "LGBP-TOP". The ensemble approach achieved the best prediction with RMSE = 9.6 and MAE = 7.6 for the reading task and 8.9, 7.2 for the free-form task, respectively. However, this method not evaluated with the test set, so the model generalisation is not fully demonstrated. Pérez *et al.* [196] proposed Gaussian "meta" approach to assess depression severity, the method aggregates the output generated from diffident modalities. These models constructed using audio, video, and silence video-audio feature extracted from AVEC-2014 dataset. The authors used Weka tool [185] to perform the "relief" feature selection approach and test the performance with the development set. The meta Gaussian regression ensemble model obtained MAE = 8.99 and RMSE = 10.82 compared to an audio-based model with MAE=9.35 and RMSE=11.9. Although this approach performed better than the baseline, it is not evaluated with the test set, and yet the audio based approach performance is somewhat sub-optimal.

Using both AVEC-2013 and AVEC-2014, Zhu *et al.* [197] applied deep convolutional neutral networks (DCNN) approach to encode facial appearance and dynamics to estimate the level of depression severity. Combining both the appearance and the dynamics modalities and obtained highest MAE and RMSE of 7.58, 9.82 for AVEC-2013, and 7.47, 9.55 for AVEC-2014 respectively. Using deep neural network to extract features from the facial raw dynamics and base the assessment on the output is a data-driven methodology, it is unclear which features characterising depression severity and how they behave compared to normal subject.

3.3.1.3 Depression evaluation using AVEC-2016 dataset

Using the AVEC-2016 (DAIC-WOZ) dataset, Yang *et al.* [198] presented multi-modal depression assessment framework consists of fusion a deep convolutional neural network (DCNN) and deep neural network (DNN) models. The proposed system used input from video, text, and audio features extracted from DAIC-WOZ dataset. Theses features include word to vector, 2D face landmarks, while audio descriptive extracted using OpenSmile [191]) software with 238 low features related to voicing and spectral energy dynamics. The fused system, both DCNN and DNN evaluated using MAE and RMSE. The performance on the development set was higher with MAE= 3.98 and RMSE= 4.65 and lowered on the test set with 5.16 and 5.97 respectively. In spite of achieving respective results, the complexity of such a model is high compare to model-based only on audio features and perform better. Williamson *et al.*[199] suggest that fusion facial, vocal articulation, and language contents modalities will improve the performance of predicting and classifying depression symptoms. The author extracted a variety of features including spectral, semantic context derived from specific questions related to depression status, filled pauses, stop word elimination, facial action unit features, the 16 delta MFCC coefficients, and loudness statistics. The Gaussian staircase model utilised to report the results, the ensemble approach achieved mean F1-score of 0.81, RMSE = 5.31, and MAE = 4.18 for the development set. This method produced acceptable results yet relying on a complex set of features evaluated with DAIC-WOZ development set only. Al Hanai *et al.* [200] utilised a neural network model known as the bi-directional Long Short-Term Memory (LSTM) to identify depression symptomology and predicting its severity.

The approach evaluated with DAIC-WOZ recordings, and a total of 279 audio dynamics applied to the audio features that provided with the dataset. The author added additional of 100 text features, both types of features used to train two LSTM models. The best performance in predicting depression scores was the fused model achieved F1=0.43, precision = 0.43, recall=0.43, MAE= 4.97, and RMSE=6.27, while best model in classification depression/non-depression was F1=0.77, precision = 0.71, recall=0.83, MAE= 5.10, and RMSE=6.37. The proposed approach has fluctuated performances between the best in predicting depression severity

Table 3.2 Summary of studies that used depressed speech Databases.

Study	Dataset	Task	Accuracy	Dev. set		Test set	
				MAE	RMSE	MAE	RMSE
Moor et al. [173]	33 subjects	Classification	96% female 91% male	n/a	n/a	n/a	n/a
low et al.[174]	139 subjects	Classification	77.8% female 74.7% male	n/a	n/a	n/a	n/a
Ooi et al.[175]	30 subjects	Classification	73 %	n/a	n/a	n/a	n/a
Cummins et al. [171]	47 subjects	Classification	77 %	n/a	n/a	n/a	n/a
Yang et al. [176]	49 subjects	Classification	69 %	n/a	n/a	n/a	n/a
Alghwinem et al. [177]	60 subjects	Classification	78.3 %	n/a	n/a	n/a	n/a
France et al. [178]	119 subjects	Classification	82 %	n/a	n/a	n/a	n/a
Meng et al. [180]	AVEC2013	Regression	n/a	6.94	8.54	8.72	10.96
Kaya et al. [184]	AVEC2013	Regression	n/a	n/a	n/a	7.84	10.22
Zhu et al. [197]	AVEC2013	Regression	n/a	n/a	n/a	7.58	9.82
Cummins et al. [182]	AVEC2013	Regression	n/a	n/a	10.44	n/a	10.17
Kachele et al. [186]	AVEC2013	Regression	n/a	7.03	8.82	8.72	10.96
Williamson et al. [179]	AVEC2013	Regression	n/a	n/a	n/a	5.57	7.4
Morales et al. [190]	AVEC2014	Regression	n/a	7.56	9.21	n/a	n/a
Zhu et al. [197]	AVEC2014	Regression	n/a	7.47	9.55	n/a	n/a
Simantirkai et al. [187]	AVEC2014	Regression	n/a	7.2	8.9	n/a	n/a
Perez et al. [196]	AVEC2014	Regression	n/a	9.35	11.9	n/a	n/a
Mitra et al. [188]	AVEC2014	Regression	n/a	5.87	7.37	n/a	n/a
Yang et al. [198]	AVEC2016	Regression	n/a	3.98	4.65	5.16	5,97
Al Hanai et al. [200]	AVEC2016	Regression	n/a	4.97	6.27	n/a	n/a
Williamson et al. [199]	AVEC2016	Regression	n/a	4.18	5.31	n/a	n/a

and best in discriminating depressed vs. no depressed subjects. Therefore the optimum model is not achieved yet. Furthermore, these results only evaluated with the development set, so it is still unclear how the system will perform if tested with the test set. Finally, Ma *et al.* [201] used combination of CNN and LSTM deep neural network approach for discrimination of depressed/non-depressed speech recording. The analysis applied to DAIC-WOZ dataset. The MFCC with 40 coefficients and their Mel-filter bank energies were extracted from the speech recordings. The author introduces a random sampling technique to train the model with balance samples; this method incorporated to overcome the problem of unbalanced classes as well as uneven speech recording length. The author claims this method will eliminate the bias associated with unbalanced classes, the best result achieved was F1= 0.52 (0.70 non-depressed), precision = 0.35(1.00) and recall = 1.00(0.54). Although this technique may reduce the bias in the performance; however, it leads to discarding information may be valuable in the evaluation. Further, the results still sub-optimal and still skewed towards depressed class (precision = 0.35) and only evaluated with the development set.

Table 3.2 provides a summery for literature regrading depression assessment studies.

3.4 Summary

In this chapter several studies have been reviewed, and the objective of these studies was to develop automatic systems to aid in detecting the early signs of dementia, cognitive related diseases and depression disorders. These studies have used variety of features including linguistics features which extracted either by using ASR or based on manual annotated texts [33, 34, 36, 150, 167, 170, 200]; such as verbs, pronouns, nouns, function words, error words, type token-ratio, total number of words, Brunet index, Honors statistics, statistics of adverbs, prepositions, etc. While, other studies used acoustic features for building their models [176, 178, 187, 201] for example, MFCCs, shimmer, jitter, voice breaks, HNR, formants, pauses, zero crossing rate, spectral energy dynamics, etc. Further, video based features were also used such as head pose, eye gaze, facial action unit, appearance descriptors, etc, and other adopted ensemble approaches that combined both audio and video features [180, 189, 194]. Identifying the most informative features, researchers used different techniques, for example, RFE, Spearman's rank correlation, CFS, statistical t-test, mRMR, mutual information, etc.

Various machine learning classification and regression algorithms were utilised by these studies for constructing the desired automated systems, such as, SVM, random forest, CNN, SGD, KNN, CNN and LSTM deep neural network, SVR, GMM regressor, etc. Also, variety of validation approaches were used such leave-one-out and k-fold cross validation to report more generalised results.

Several limitations were found in these studies, for example, validated using few samples [53, 139, 146, 156] or with unbalanced dataset [138, 143, 161, 162]. These limitations and gaps will be discussed in following chapters 4, 5, 6, 7 and 8 when the proposed systems introduced and the results were compared.

Chapter 4

Detecting early signs of dementia

4.1 Introduction

This chapter demonstrates the feasibility of developing a simple and robust automatic system based solely on acoustic features to identify Alzheimer's disease (AD) with the objective of ultimately developing a low-cost home monitoring system for detecting early signs of AD. As mentioned before in chapter 2 section 2.1.3, there is no powerful tool that gives a reliable diagnosis of dementia; rather, the patient has to go through a series of cognitive tests conducted by a professional neurologist for assessments. This process can be very challenging for the patient and involves a certain amount of anxiety and stress. Especially in the case of the early stage detection, complementary tests include the analysis of samples of cerebrospinal fluid taken from the brain and a magnetic resonance brain imaging test [5]. Such methods are invasive, bring discomfort to the patients, are relatively costly and require a significant amount of effort and time. Finding lightweight, noninvasive diagnostic and/or screening tools, that can be used in the comfort of peoples' homes and inform this process, is therefore of interest. This could be in the form of wearable sensors or incorporated in existing intelligent home technology. This work describes a relatively simple audio-based tool for detecting biomarkers of dementia in a person's speech, this work will be evaluated using DementiaBank dataset [147].

The rest of the chapter is organised as follow: section 4.2 describes the DementiaBank corpus and provide details about the patients, demographic information and performed task. Section 4.3 is the methodology section that introduces the proposed system pipeline and provides a description for each component. After that, the performance of the proposed system presented in the results section 4.4. Section 4.5 is the discussion section, in which the results were discussed and compared to other studies from the literature. The last section 4.6 s the summary and conclusions.

4.2 Dataset

The DementiaBank data set is a free access large existing database for Alzheimer's and related dementia diseases collected during longitudinal study conducted by the University of Pittsburgh School of Medicine and as part of Alzheimer Research program [202]. A verbal description of the Boston Cookie Theft picture (see Fig 4.1) was recorded from people with different types of dementia (such as probable or possible AD, vascular dementia,etc.) with an age span from 49 to 90 years as well as from elderly HC subjects with an age range from 46 to 81 years[151]. The speech samples were transcribed using the CHAT transcription format Mac Whinney [203].

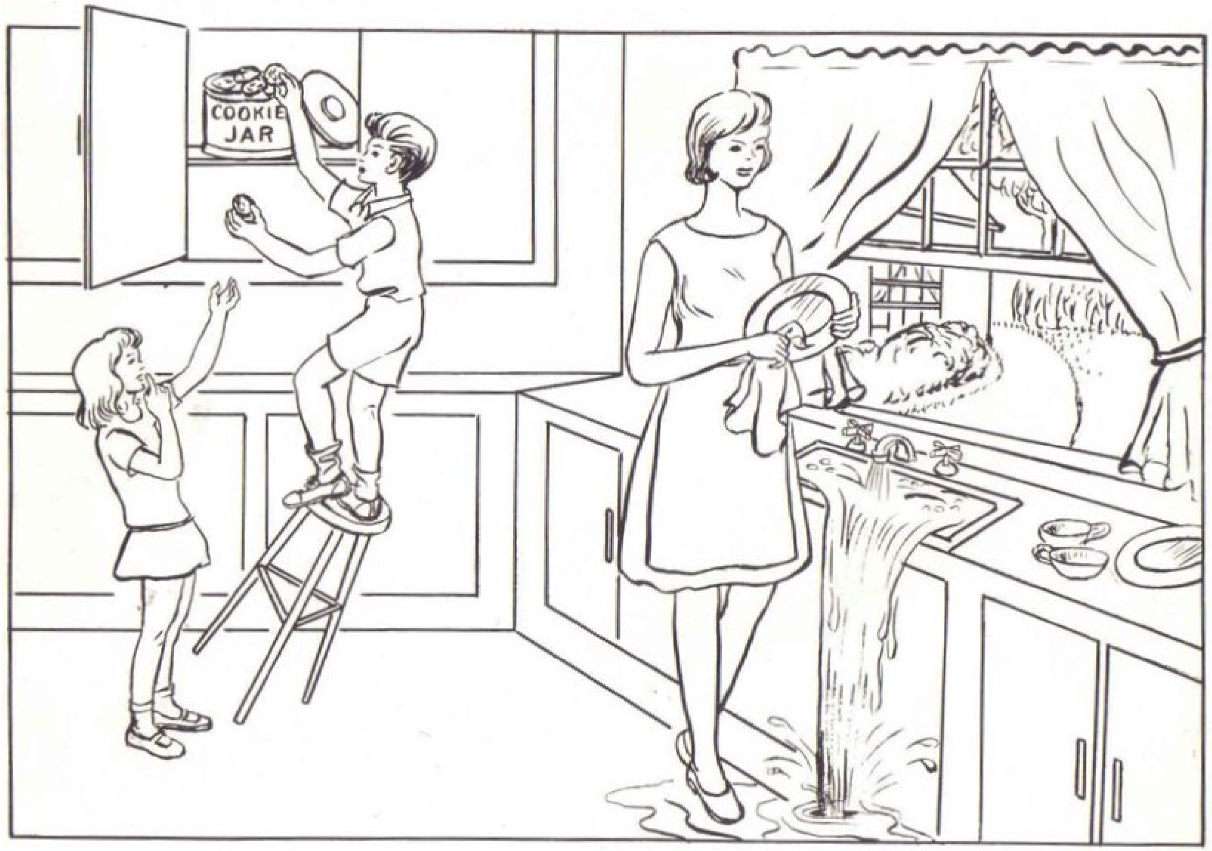


Figure 4.1 The "Cookie Theft Picture" from the Boston Diagnostic Aphasia Examination [2].

Table 4.1 DementiaBank data set demographic information.

Group	HC	AD
Diagnostic lables	Healthy	probable and possible AD
No. participants	97	167
No. of samples used	233	240
Age	64.5	71.8
Education (years)	14	12
Sex(M/F)	39/58	56/111
MMSE	29.1	18.7

The Cookie Theft picture which it is a part of the Boston Diagnostic Aphasia Examination [2], and it is mainly utilised for capturing narrative speech from the speakers in order to diagnose the different types of language and communication disorders. In the picture the woman is drying plates and not paying attention to the overflowing sink, also, there are a boy and a girl trying to steal cookies from a cookie jar placed in a kitchen cupboard. The boy is using a stool and about to fall down. During the interviews, patients were given the picture and were told to discuss everything they could see happening in the picture. The descriptions were recorded on a yearly basis, the first visit date varied between subjects from (1983-1988), the last 7^{th} visit was recorded in 1996 by very few participants. Table 4.1 shows the demographic information for both AD and HC groups used in this experiment. The AD group contains 240 speech recordings for patients having either probable AD or possible AD diagnosis, while HC group have 233 speech samples.

4.3 Proposed system

The general pipline for the proposed system is shown in Fig 4.2, and it consists of preprocessing, feature extraction, and machine learning units. The following sections will provide details regarding each components.

4.3.1 Pre-Processing

The first step of the pre-processing is background noise reduction, the speech recordings of DementiaBank dataset contains a high level of background noise. Effective de-noising is important to enable accurate features extraction. Therefore, the spectral noise gating method applied to eliminate the noise without sacrificing the quality of the desired speech recording. Fig 4.3 shows

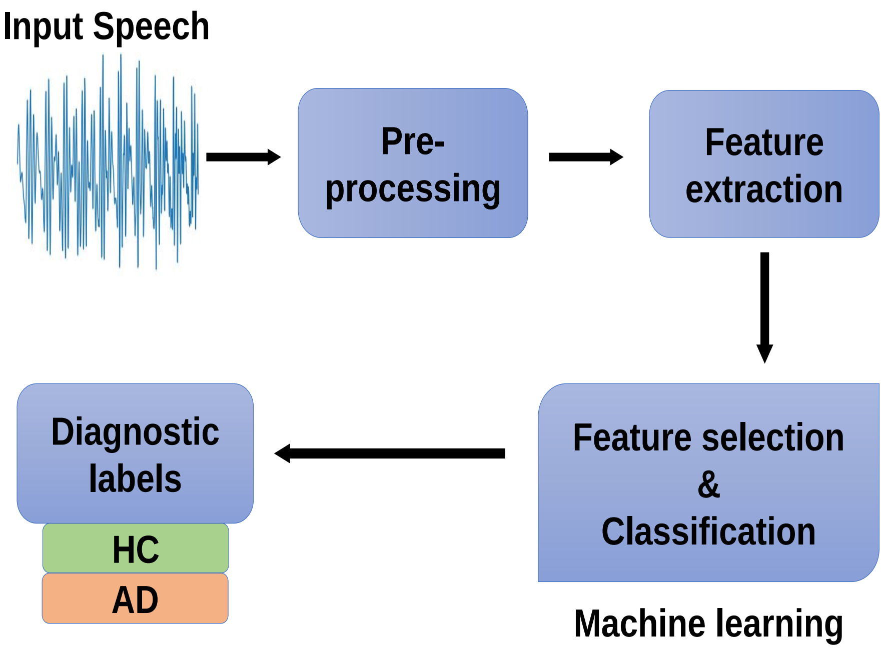


Figure 4.2 Proposed dementia detection system.

the effect of the denoising process. This process carried out using the Audacity software [204]. The spectral noise gating works by defining a threshold (power level); the signal passes when the power is higher (i.e., utterance) or attenuates (i.e., background noise) when the power is lower than the threshold, however this process was done manually to identify the noise threshold and for each speech sample because the noise profile varies between the recordings. The quality of the denoised recordings was inspected to ensure the readability of the audio recordings. This method showed an efficient suppression of the noise that helped in the voice activity detection task and eventually improving the system performance.

4.3.2 Features extraction

The focus of the study was on extracting only the acoustic features and investigating the effectiveness of these features in detecting dementia at its early stages. This eliminates the need for manually transcribed files or indeed the problems around achieving reliable speech recognition results, especially in challenging far-field acoustic conditions.

Table 4.2 summarises the 263 features extracted in this chapter as as follows:

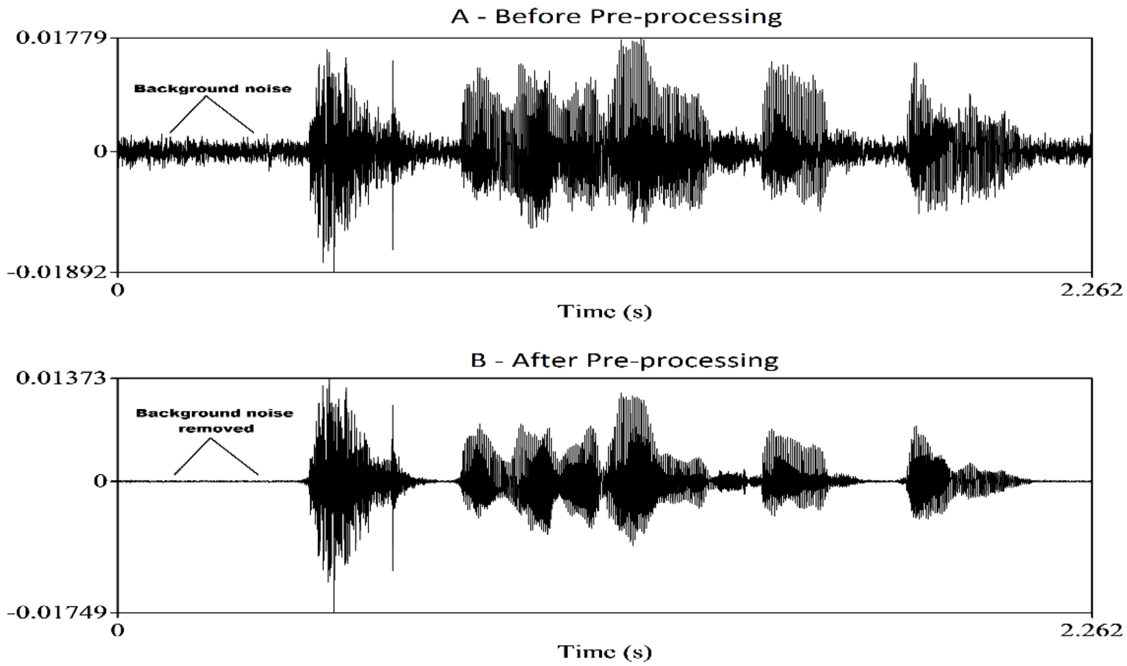


Figure 4.3 Speech sample before (A) and after (B) the pre-processing step

4.3.2.1 Phonation and voice quality features

Phonation and voice quality features were included because previous studies found them to be predictive of AD diagnoses. Meilán *et al.*[52] and Lopez-de-Ipena *et al.*[205] achieved accuracies of 84.8% and 96.9% respectively when classifing between AD and HC subjects using these acoustic characteristics. This group of features includes the fundamental frequency (F0) and its related variances (shimmer and jitter). The F0 is a measurement of vocal fold oscillations [206] that are known to be nearly periodic in healthy voices, but less so in voice pathologies [207, 208]. Jitter$_{(local)}$ describes the frequency alteration from cycle to cycle, while the shimmer $_{(local)}$ measures the amplitude fluctuations of the consecutive cycles. [159] provides further details of these parameters.

The voice quality parameters included; the harmonic-to-noise ratio (HNR) which measures how much energy there is in the periodic part of the signal compared to its non-periodic part; and noise-to-harmonic ratio (NHR) which measures the amplitude of the noise generated due to incomplete closure of the vocal folds during the production of the speech relative to tonal components [209]. Additional features included the; number of voice breaks (distances between pulses greater than 16 milliseconds); degree of voice breaks (ratio of the breaks' total

Table 4.2 Summary of all features extracted in this chapter.

#	Features set	Description
1	Phonation and voice quality	The total verbal time
		Pitch variation features (mean, median, STD, Min and Max)
		Mean periods and STD periods
		Fraction of locally unvoiced frames and degree of voice breaks
		Jitter: (local, local-absolute, the relative average perturbation (rap), five-point perturbation quotient (ppq5) and the average absolute difference (ddp).
		Shimmer: (local, local-dB, three-point amplitude perturbation (apq3), five-point amplitude perturbation quotient (apq5), eleven-point amplitude perturbation quotient (apq11) and the average absolute difference (dda).
		Mean of autocorrelation
		Mean noise-to-harmonics ratio
		Mean harmonics-to-noise ratio
2	Speech and silent statistics	Max, mean, median and STD of speech segment length >=0.4 sec
		No. of pauses (pause length of >=1ms are considered)
		Total speech & silent durations for the segments >= 0.4 sec
		Max, mean, median and STD of silent segment length >=0.4 sec
		Total silent length >=0.4 sec. including the pauses
		Number of speech and silent segments>=0.4 sec.
		Mean and STD of pauses and total duration of the pauses
3	Spectral features	26 Spectral centroid coefficients
		26 Filter bank energy coefficients
		First 42 MFCC coefficients and their skewness, kurtosis, mean with kurtosis and skewness of the mean
	Total	**263**

duration to the total duration of the analysed signal), these features were extracted using Praat tool [159]). These features were inspired by the work presented by Meilán *et al.* [52]. The authors showed that the same features proven to be useful in identifying AD patients from healthy control people.

4.3.2.2 Speech and silent statistics

The second group of features were derived by applying machine classification algorithms to identify speech/non-speech segments. This is done by windowing the audio files into 25ms frames with 10ms overlapping window. For each frame apply the short time energy, zero crossing rate and the correlation coefficients. These three measures with labelled frames are used to train and build a voice activity detection (VAD) classifier using predefined frame samples randomly selected from the data. Next, the VAD was used to label each frame for the rest of the audio files. The results from the VAD classifier gives the duration statistics for speech/silent regions with the amount of pauses presented in the recordings. These features were intensively used in previous studies and proven to be of great importance in detecting dementia related signs [54, 131, 144, 161, 210].

4.3.2.3 Spectral features

Speech production involves the movement of articulators including the tongue, jaws, lips, and other speech organs. The position of the tongue plays a key role in creating resonances in the mouth. Although speech articulation is relatively preserved in the commonest types of AD, it is conceivable that AD could have measurable effects on the coordinated activity of speech articulators and thereby spectral features. Mel frequency cepstral coefficients (MFCCs) [211] were extracted to capture the spectral content of the speech signal. The hypothesise is that patients in the AD group might be characterised by lower spectral coefficients valued than those in the HC group. The MFCCs aims to compute the energy variations between frequency bands of a speech signal. MFCCs have become widely used in speaker verification, speech recognition and for the extraction of paralinguistic information since they were proposed by Davis and Mermelstein back in 1980. The MFCC computed following the method explained in [212], in which, each 25ms frame is converted into the frequency domain using the fast Fourier transform (FFT) before the power spectrum is estimated by taking the absolute value of the complex FFT and squaring the result. Next, the Mel triangular filter-banks are applied and calculated by converting the frequencies into the Mel scale and summing the energy for each filter. By taking the logarithm of all filter-bank energies, result in obtaining second set of features, namely the logarithmic energy of the Mel filters (Fbanks; 26 features).

The last set of SF features were the spectral sub-band centroids (SSCs). SSCs aims to locate the spectrum centre of mass and found to be valuable in measuring the cognitive load le *et al.* [213]. Therefore SSCs could be useful in this study, and a total of the first 26 coefficients were included, SSCs are extracted by dividing the energy in each filter-bank (i.e., 26 filter-banks as estimated previously) by the total energies of all filter-banks [214, 215]. The spectral features were only applied to the participants utterances in the recordings (i.e., excluding the silences and the instructor utterances). The final representation of these features includes: the first 42 MFCC coefficients and their skewness, kurtosis, means and kurtosis and skewness of the means) in addition to the first 26 coefficients for Fbank and SSCs.

4.3.3 Feature selection

Features must be explored and examined in order to achieve maximum performance. Feature selection techniques identifies the most significant ones and discard redundant, and the ones that reduce the model performance. Feature selection may also decrease the risk of over-fitting [216, 217]. Three approaches were used to determine the importance of the features. The first is known as the wrapper. The wrapper method was used based on an SVM evaluator known as the recursive feature elimination (RFE) technique [218], in which, the features are eliminated sequentially and the model performance estimated each time until all features have been excluded. The feature that has the maximum negative impact on the result is considered to be the most important one. Likewise, the rest of the features are then ranked. The second approach is known as the ensemble based on features importance technique used in the case of tree-based classifiers (i.e., random forest, Adaboost, and Bagging based on the tree). The feature importance approach gives a score that indicates how valuable each feature was in the construction of the boosted decision trees within the model. The feature that frequently used to make critical decisions with decision trees will poses higher relative importance. This importance is computed explicitly for each feature in the dataset, allowing attributes to be ranked and compared to each other. The importance is calculated for a single decision tree by the amount that each feature split point improves the performance measure, weighted by the number of observations the node is accountable for. The feature importances are then averaged over all of the decision trees inside the model.

The last method is to conduct statistical analysis and examine all variables. The null hypothesis assumes there is no significant difference between the means of the two classes for a specific variable. The assumption is that the features that reject the null hypothesis will be selected and ignore those features that accept it. The SPSS software [219] was utilised to perform Mann-Whitney u-test appropriate for non-parametric data because the Shapiro-Wilk normality test [220] suggested that these features were not normally distributed. Table 4.3 lists the top ranked features using the wrapper and the statistical U-test (0.05 significance level) at 95% confidence interval.

Table 4.3 Top ranked features using the wrapper and the statistical U-test. U represents Mann-Whitney u-test and P^* << 0.05

#	Classifier-wrapper rank	Weight	Statistical rank	U	P
1.	Mean-MFCC2	82.241	Mean-MFCC20	16122	0.00*
2.	Kurtosis -MFCC30	81.606	Skewness-MFCC14	16843	0.00*
3.	Mean-MFCC30	81.606	Mean silent seg.	17223	0.00*
4.	Skewness- MFCC2	80.972	Mean-Fbank10	17304	0.00*
5.	Mean-MFCC16	80.126	Mean-MFCC2	17337	0.00*
6.	Mean-Fbank 22	79.069	Mean-MFCC14	17657	0.00*
7.	Spectral centroid -C14	79.069	STD of silent seg.	17772	0.00*
8.	Mean-MFCC30	77.801	Mean pauses time	17772	0.00*
9.	Kurtosis -MFCC16	77.589	Kurtosis -MFCC30	17919	0.00*
10.	Mean-Fbank 2	77.589	Spectral centroid -C14	18235	0.00*
11.	Mean-Fbank 24	77.167	Skewness- MFCC2	18576	0.00*
12.	Mean-MFCC1	76.532	Mean-Fbank 22	18715	0.00*
13.	Mean-Fbank15	76.052	Skewness- MFCC32	18735	0.00*
14.	Kurtosis -MFCC2	73.995	Spectral centroid -C20	18929	0.00*
15.	Mean-Fbank 20	72.304	Fraction of locally unvoiced frames	18932	0.00*
16.	Mean-Fbank 13	65.961	Skewness- MFCC24	19004	0.00*
17.	No. of silent segments	61.522	Total silent length	19102	0.00*
18.	Fraction of locally unvoiced frames	59.830	STD of pauses	19102	0.00*
19.	Minimum silent segments length	57.928	Mean-Fbank2	19140	0.00*
20.	Median pitch	49.48	Kurtosis -MFCC16	19654	0.00*
21.	-	-	Mean-Fbank24	19900	0.00*
22.	-	-	No.of pauses >= 1sec	22134	0.00*
23.	-	-	Mean silent seg.	24044	0.0082
24.	-	-	Total No. of silent seg. >= 0.4sec.	24056	0.0086
25.	-	-	Total Pause time	24056	0.0086
26.	-	-	Median pitch	24305	0.014

4.3.4 Validation Scheme

In the machine learning community, cross-validation is widely used to ensure an effective method of model selection to achieve, a robust performance evaluation and prevent over-fitting [221]. The K-fold cross-validation with k = 10, was utilised to partition the data into ten equal parts called "folds". The model was trained using nine out of ten folds and tested with the remaining 10th fold. This step was repeated k=10 times until all folds had been used in the training and testing process. This however, did not generate the validation set directly. Instead, the nested k-fold cross-validation method adopted, which uses two k-fold loops namely the outer and the inner loops. The outer loop generates the testing (1/10 data) and the training (9/10 data) folds, while the inner loop takes all the training folds combined (coming from the outer loop) and generates the validation and training folds. Feature selection and the model's hyper-parameter tuning were explored and the model with the best features and best parameters was tested using the test folds. This process runs through all the loops and the final model result is reported as the average of the best model's scores across the outer test folds. Importantly, each fold generated had to contain a balanced number of samples between the two classes for the nested k-folds cross-validation in order not to skew the output towards one class. The Stratified-KFold from Scikit learn library was used to perform this task [218]. The StratifiedKFold which is a special type of kfolds that maintain approximately the same number of samples for each targeted class in the generated folds(i.e. the training, validation and test sets). Although this is a minor issue in this experiment due to the fact that the number of samples in both AD and HC classes does not differ significantly (AD n=240 and HC n=233), it was used to ensure the best possible approach for this kind of problems.

One important step also used and before training any models, which is the feature normalisation. This preprocessing step was executed at the training phase (training data) and excluded from the validation and test phases. Different methods can be used for feature normalisation, and in this work, a method expressed in equation Eq 6.1 was used, which is known as the standard scalar, and for a training sample x is given by:

$$Stand(X) = \frac{x - \mu}{\sigma} \tag{4.1}$$

Where μ and σ is the mean and standard deviation of the training samples respectively.

4.4 Results

The results obtained using the capability and accuracy of the automated machine learning algorithms, which helped to evaluate the potential of the acoustic features to distinguish between AD patients and HC subjects. Four different machine learning classification algorithms were used including: Bayesian Networks (BN), Trees-Random Forest (RF), AdaboostM1 (AB) and Meta- Bagging (MB). These classifiers were utilised to achieve the final results in four different configurations resulting from using pre-processing or not, and using the full (263) or the reduced features sets, also the sensitivity and specificity were estimated for the highest scores achieved (i.e., for the 2^{nd} and the 4^{th} configurations). Table 4.4 lists the accuracies obtained for

Table 4.4 Shows the performance under different running configurations.

#	Machine Learning Algorithm	1^{st} Configuration: 263 features	2^{nd} Configuration: Top 22 features	3^{rd} Configuration: Top 26 based on U-test	4^{th} Configuration:Pre-processing with top 20 features
		Accuracy %	Accuracy %	Accuracy %	Accuracy %
1.	Bayes Net	80.3	91.75	90.06	94.71
2.	Meta-Bagging	81.8	92.38	88.7	92.6
3.	Random forest	82.8	91.96	91.75	92.8
4.	AdaBoost M1	79.9	85.83	86.6	91.75

the four different configurations. The highest classification accuracy achieved was 94.71% using the (BN) classifier, running under the fourth configuration followed by configuration three with 93.66% using (BN) classifier, while configurations two and one score 92.38% and 82.8 % using (MB) and (RF) classifiers respectively. By adopting a pre-processing step and extracting fewer, better quality features for the classifiers, the highest accuracy was achieved. The sensitivity and specificity for the 2nd configuration was 92.00%. Only 19 patients from 240 and 17 HC subjects from 233 were incorrectly classified, but when comparing with the 4th configuration, only 7 AD patients were incorrectly classified making the sensitivity level at 97.00%. However the specificity of the 4th configuration was slightly reduced to 91.00% (only 21 HC were misclassified)

4.5 Discussion

Speech and language impairment serves as strong evidence for Alzheimer's disease detection [170]. For the same data set (based on short speech recordings from a picture description task.), but using only acoustic features, higher accuracy results were obtained, in distinguishing between HC subjects and AD patients, than those reported in the most recent state of the art [34].

The results show that acoustic features carry valuable information. The Median pitch and the fraction of locally unvoiced frames were the only informative variables from the first group of features. This is on the contrary to what has been reported in the previous study by Melan *et al.* [52]. Meilan *et al.* shows that HNR, shimmer and jitter were significant in identifying AD patients, this contrast between the findings might be due to the differences between the performed tasks (picture description vs reading) or due to the recording environment and devices. Pauses and number of silent segments are more prevalent in AD patients as they tend to shorten the speech segments in contrast to HC subjects. This is because the AD patients, and when cognitive load increased, either they lost interest in the subject or they find that talking requires much effort and concentration.

The statistical analysis also indicates those variables have significant mean difference between AD and HC classes, for example, the total verbal time (66.0 sec for AD compared to HC 55.9 sec), the silent (AD= 38.5 sec vs HC =25.4sec), also, the number, mean and total pauses, (AD = 21.28 sec, 1.88, 37.7 sec compared to 18.46 sec, 1.39, 24.1 sec for HC) respectively, on the other hand, HC subjects produced more utterances, longer speech segments, have higher utterance ratio to the total time of 30.4, 4.2 sec, 0.5 respectively compared to AD of 27.5, 3.6 sec, 0.39 respectively, these findings were similar to Singh *et al.* [53] and Roark *et al.* [54] studies, both reported that the average time of both pauses and speech are useful in discriminating healthy control participants from mild cognitive impairment and AD patients.

Other notable features were the MFCC dynamics, although they are well-known as standards in speech recognition systems, capture important separation between the two groups as they relate to the articulators (libs and tongue) control ability, that is decreased in AD patients [222], both Fraser *et al.* and Yancheva *et al.* [34, 170] reported that skewness and kurtosis of MFCC coefficients were included during the feature selection process; thus they contain valuable information for identifying AD. In addition to MFCCs, both filter bank and spectral based features were informative, therefore selected in both feature selection methods.

The noise reduction procedure played a significant role in improving the accuracy; first, it helped to identify the utterances more efficiently compared to the unprocessed data because the proposed method is not relying on the transcripts to determine the participant's turns. Secondly, it helped to extract more accurate and robust features; for example, the HNR and NHR features aim to measure how much information and noise ratios presented in the voice segments. As a result, models built with de-noised data achieved higher accuracy.

Comparing this system to other modalities from the literature which used the same dataset, Orimaye *et al.* [33] used linguistic features extracted from the manual annotated text files, and they reported 74% classification accuracy using a support vector machine (SVM) classifier with 10 folds cross-validation, later Orimaye *et al.* [148], used only 99 samples from the first visit from both AD and HC participants, and with use of n-gram features with 1000 variables, they achieved an area under receiver operating characteristics of 0.93. Other study by Sirts *et al.* [35] which also used linguistic elements including the propositional idea and semantic idea densities features, the study achieved F1-score of 75%. Further, Ammar *et al.* [150] reported SVM precision of 0.79% using syntactic, semantic and pragmatic features. Another linguistic approach developed by Klumpp *et al.* [152], their model used bag of words based features to construct an artificial neural network model. The authors reported classification score of 84.4%. Whereas Mirheidari *et al.* [154] used ward vector representations derived from using first manual annotated files and second based on automatic speech recognition (ASR) annotation, and the best result was based on the first model with accuracy of 75.6% compared to 62.3% form using the ASR. Furthermore, a study conducted by Fraser *et al.* [34], in which the authors extracted a set of 370 acoustic, lexical and semantic features, the best classification accuracy was 81% in distinguishing between HC and AD groups. In all of these studies, the proposed system presented in this chapter, outperformed their modalites and by using simpler and automatic approach. Also, these studies were based on the manual annotations, and such systems may not be applicable for real-time scenario. Furthermore, Mirheidari *et al.* [154] showed that when ASR used in an automated screening system, the performance reduced, while other studies did not report how their models will behave when ASR replace the manually annotated texts.

Finally, the only study have been found which solely based on acoustic features is presented by Luz *et al.* [223]. Luz *et al.* aimed for simple design to longitudinal monitor of AD progression. The authors used 214 samples from AD and 184 from HC group. They extract paralinguistic features including speech rate, pauses and vocalisation timing variables. Luz *et al.* used Bayesian classifier to report the results with 10-fold cross validation method. The authors reported an overall accuracy of 68% without using feature selection technique which might improved their results.

4.6 Summary

In this study, the experimental results demonstrated the efficacy of set of acoustic features extracted directly from speech recordings for individuals preforming short vocal task. Higher accuracy results were obtained in distinguishing between HC subjects and AD patients, than those reported in the most recent state of the art [34].

Furthermore, the acoustic features derived automatically from the speech recordings without the addition of any lexical or syntactic features that rely on complex speech recognition technology as in [224]. This chapter proposed a simple high accuracy automated method that can be used in the clinic and/or at home to guide the diagnosing and/or screening of dementia by using just speech. The proposed method also robust and very capable of identifying dementia patients from healthy individuals even in the presence of significant background noise. These facts support the proposition for using only acoustic features as an objective tools for automatic detection and/or screening of AD at a low cost and within the home environment.

Chapter 5

Longitudinal detecting and predicting the severity of AD

5.1 Introduction

The current clinical procedure (as mentioned previously in chapter 2 section 2.1.3) for diagnosing dementia is based on cognitive tests of which the Mini Mental State Examination (MMSE) test is the most commonly used. It measures the severity of cognitive impairment in five categories: attention, registration, language, orientation and memory. The MMSE test consist of 11 questions with a top score of 30 points (considered a normal cognitive state) and where a score of 0 reflecting a major cognitive decline [61]. Although this method is widely used as a standard clinical screening procedure, it is performed using pen and paper and the test also relies on the presence of experienced neurologists. Using the MMSE unified cognitive scale in automatic screening models would be highly useful for clinicians, in terms of providing faster assessment and supporting consistent diagnosis by the non-experts.

Based on the evidence found in studies [146, 225] which reported a decline in speech and language characteristics due to AD. Therefore the work in this chapter hypothesis that using features extracted from speech recordings of patients suffer from AD might be used to develop a system is capable of predicting scores similar to MMSE. This work proposing an automatic speech analysis method applied in the context of AD screening and diagnosis. The proposed regression model will predict the MMSE scores from speech recordings both within a single clinical visit and for future recordings. This would potentially mean that following an initial MMSE test with associated speech recording, for subsequent assessments, the MMSE score could be predicted from a new speech sample alone. That is, the patient can avoid having to do the potentially stressful MMSE tests. Ultimately, this might enable clinicians to monitor the deterioration of a patient in between scheduled visits to a clinical setting as a speech recording (done at home) could provide an estimated MMSE score.

This work also extend the functionality for the system that introduced in the previous chapter 4, and the new model will classify patients from their speech recordings as Alzheimer's disease (AD), healthy control (HC) individuals or patients with Mild Cognitive Impairment (MCI). A key novelty of this work is that only acoustic features (including a new class of spectral features) are used, which has the benefit of avoiding the cost and the complexity of employing automatic speech recognition technology and higher level spoken language understanding.

The rest of the chapter is organised as follow: Section 5.2 describes the longitudinal dataset derived from DementiaBank corpus. Section 5.3 illustrates the methodology and describe the proposed system and its units. While section 5.4 presents the results for a number of scenarios developed to demonstrate the efficacy of this system. Section 5.5 shows the discussion, while section 5.6 summarises this work.

5.2 Dataset

The DementiaBank data set is already described in the previous chapter 4, section 4.2. However, in this work, only longitudinal evaluation will be performed, Table 5.1 shows the demographic information for only participants who completed three visits. Only 16 AD patients successfully follow up three visits; thus the proposed dementia severity evaluation system will be evaluated using these AD samples in the three visits. Further, in the same table, a detailed for other longitudinal groups, namely HC and MCI, were listed. The three groups will be used to develop a classification system, and this system will be evaluated for each visit samples.

5.3 Proposed system

The same pipline introduced in chapter 4 section 4.3 will be used in this work, however, this time two models will be evaluated using machine learning classification and regression algorithms,

Table 5.1 Three visits dataset demographic information.

Variables	All subjects	AD	MCI	HC
	(N=64)	(N=16)	(N=6)	(N=42)
Female	37	8	2	27
Male	27	8	4	15
Age (Mean)	66.76	73.93	70.8	63.45
(STD)	8.8	9.27	4.7	7.0
Education	14.6	23.25	28.0	29.16
(years)	3.17	3.5	1.15	1.02
MMSE 1	27.57	22.62	28.0	29.11
	3.2	3.4	1.15	0.98
MMSE 2	27.0	20.81	27.3	29.11
	3.7	3.9	1.79	0.98
MMSE 3	26.0	17.43	27.8	29.12
	5.8	4.6	1.93	1.34

thus the updates pipline illustrated in Fig 5.1. This system have the same preprocessing and feature extraction units as in the system introduced in chapter 4 (Fig 4.2). The following sections will explain each unit.

5.3.1 Pre-Processing

This unit will have the same functionality as described previously in chapter 4 section 4.3.1.

5.3.2 Features extraction

This work will use the same phonation and voice quality features as described in chapter 4 section 4.3.2. However, the spectral features expanded to include new features and as follows:

5.3.2.1 Statistical descriptive features

This group of features extracted by applying the common statistic functions to the spectral feature group and as follows: For the spectral features (SF) which includes 42 values of the Mel Frequency Cepstral Coefficients (MFCC), the logarithmic measures for the first 26 filter bank energies (FBANK energy coefficients) and the Spectral Sub-band Centroid (SSC) with 26 coefficients. A total of 658 (noted as SSF1) features were extracted including seven statistics measures: the standard deviation (STD), mean, max, min, median, skewness, and kurtosis applied to (SF). Also, second level of the same seven statistics (excluding the median) were applied to the SSF1, which added extra 126 features. As a result the total feature vector increased

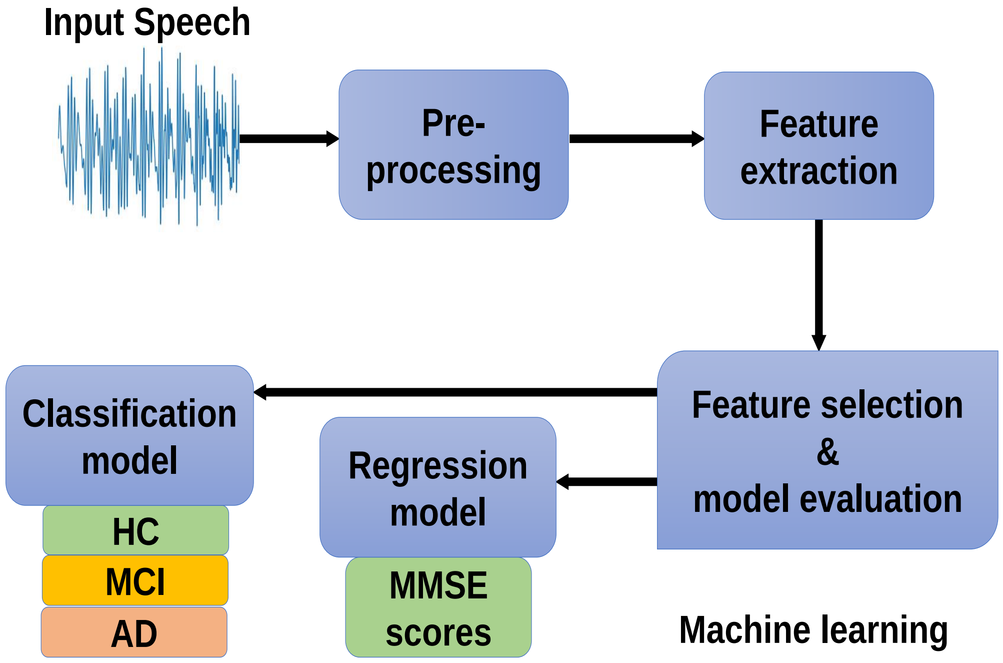


Figure 5.1 Proposed system.

to 812.

Table 5.2 List of all acoustic features used in the proposed system.

Features	Type	Number of features
Fundamental frequency (F0) related measures (median, mean, STD, min and max)	Phonation and voice quality	5
Harmonic-to-noise ratio (HNR)	Phonation and voice quality	1
Number of pulses	Phonation and voice quality	1
Number, mean and STD of periods	Phonation and voice quality	3
Noise-to-harmonic ratio (NHR)	Phonation and voice quality	1
Shimmer scales	Phonation and voice quality	6
Jitter scales	Phonation and voice quality)	5
Autocorrelation	Phonation and voice quality	1
Fraction of locally unvoiced frames	Phonation and voice quality	1
Number of voice breaks	Phonation and voice quality	1
Degree of voice breaks	Phonation and voice quality	1
Number of responses	Speech and silence	1
Average responses time	Speech and silence	1
Mel frequency cepstral coefficients (MFCC)	Spectral features: 42 features	(extended to 336)
Filter bank energy coefficient (Fbank)	Spectral features: 26 features	(extended to 224)
Spectral subband centroid (SSC)	Spectral features: 26 features	(extended to 224)
Total		812

5.3.3 Feature selection

In this work a similar approach adopted, which was described earlier in chapter 4 section 4.3.3, however, this time the the Support Vector Machine (SVM) was used as an evaluator classifier because it performed well in selecting the significant features, this wrapper method implemented using Recursive Feature Elimination (RFE) Scikit-learn libraries [218].

5.3.4 Validation scheme

Three scenarios are considered in the classification of the three groups (AD, MCI and HC), namely: (HC vs AD), (HC vs MCI) and (AD vs MCI). It is more realistic to run the three classification scenarios separately for each one of the three visits. To do that, two machine learning classification algorithms were used for the evaluation: the SVM and linear via the Stochastic Gradient Descent (SGD) optimisation method, and both running in nested leave-one-out cross-validation (LOOCV) approach. This approach means that there are two LOOCV loops. The first loop (outer loop) will divide the dataset into n folds and reserve one sample for the test and trains the model with (n-1) samples. The second loop (inner loop) will partition the (n-1 samples generated from the outer loop) again into training set having ((n-1) -1) samples and the remaining one sample as a validation set. Feature selection is deployed during the inner LOOCV loop, and the nominated features will be evaluated with the validation sample. The classifier hyperparameters will be selected during the inner loop. The best model with the best parameters will be tested in the remaining sample from the outer loop. The final classification accuracy represents the average scores for all outer testing folds. The LOOCV was selected instead of the K-fold because the number of samples is few 64 per visit compared to 473 for work performed in chapter 4. The classification procedure is done separately for each visit. In each visit, there is only one speech recording per patient, for example in visit one, there are 42 HC subjects having 42 speech recording, 16 AD patients provided 16 speech recordings and 6 MCI patients having six recordings. Therefore LOOCV will not deal with multiple samples for each participant when generating the training and testing folds.

The main objective of this work is to predict the MMSE clinical scores for the AD patients only, the reason behind that is the AD patient's are more likely to experience rapid decline in cognitive level (mild, moderate and sever AD status). This decline is expected to be in a shorter time compared to HC and MCI individuals. That is seen very clearly from the three visit's MMSE scores. In this task, the random forest regression algorithm was selected and running under the same procedure as for the classification task (i.e. the feature selection and LOOCV method). Since three visits for the AD group were derived from the DementiaBank, therefore multiple scenarios were designed: predicting the MMSE within each visit (MMSE1, MMSE2 and MMSE3 respectively), and in addition, also to predict the MMSE evaluation across the visits using the model built by visit 1 samples to predict the MMSE score for visit 2, and likewise predicting from visit 2 to visit 3. Finally, to investigate the effect of combining visit 1 and 2 speech samples to predict MMSE scores for visit 3. The last three configurations will test the system capability in predicting a future MMSE scores. This approach will benefit the clinicians and healthcare providers terms of cost saving and reduce the need for the experts that perform such longitudinal investigation. Since there were three visits available MMSE1, MMSE2 and MMSE3. The difference between the scores will be used to assign the regression labels, also these values will be adjusted by the time which the MMSE evaluation is conducted, the following equations (5.1) and (5.2) will be used to produce the future MMSE scores and as listed in Table 5.3:

$$\Delta\ MMSE = Current\ MMSE - Next\ MMSE \tag{5.1}$$

$$Future\ MMSE = \ Current\ MMSE - \left(\Delta\ MMSE * \frac{T}{365}\right) \tag{5.2}$$

Where *T* represent the period between the current MMSE visit date and the next MMSE visit date, and *T* is estimated in days. When visit dates were not specified *T* will be assigned the value of 365 days (the common period between the visits as observed in the dataset's demographic sheet).

5.4 Results

The results reported in terms of the average classification accuracy for all visits for all the classification scenarios. Tables 5.4 and 5.5 summarises the results. To create balance between the classes, the Synthetic Minority Over-sampling Technique (SMOT) [226] was utilised, and

Table 5.3 Three visits dataset MMSE and their future MMSE scores estimation

ID	MMSE1	Visit1	Future MMSE2	MMSE2	Visit2	future MMSE3	MMSE3	Visit3
010	20	09/08/1983	21.09	21	10/09/1984	26.08	26	16/09/1985
051	26	30/11/1983	23.07	23	20/11/1984	18.53	19	02/01/1986
057	27	07/12/1983	24.49	24	07/10/1984	20.81	13	21/01/1985
058	23	04/01/1984	22.02	22	26/12/1984	17.67	17	07/11/1985
076	25	24/02/1984	21.19	20	28/11/1984	20.00	20	N/A
091	19	15/03/1984	17.14	17	17/02/1985	17.00	17	09/05/1986
094	27	13/03/1984	23.01	24	11/07/1985	24.00	24	N/A
134	24	10/08/1984	22.74	23	14/11/1985	6.00	6	N/A
157	19	25/07/1984	16.64	17	29/09/1985	9.00	9	N/A
164	24	13/07/1984	24.00	24	06/08/1985	18.90	19	13/08/1986
181	20	14/09/1984	16.46	17	19/11/1985	17.00	17	N/A
212	25	05/02/1985	18.54	19	05/03/1986	10.86	17	29/03/1990
213	16	04/12/1984	14.98	15	11/12/1985	15.00	15	N/A
252	22	24/04/1985	23.15	23	19/06/1986	16.95	17	22/06/1987
270	23	18/06/1985	21.95	22	06/07/1986	19.65	20	08/09/1987
282	22	12/07/1985	22.00	22	29/09/1986	22.96	23	16/09/1987

similarly all the classification scenarios were evaluated and compared to the results from the original unbalance samples.

SMOT use K-Nearest Neighbour algorithm to create samples for the minority class. Table 5.5 showing the performance of the classifiers and the three balanced group samples when using SMOT. Classifying between AD and HC (scenario 1) the average classification accuracies were 93.4%, 97.0% and 91.4% respectively. In comparison, the classification accuracies for the unbalanced data were all lower at 88.5%, 94.8% and 85.3% for SVM which was all equal or higher than using SGD. The accuracy goes up for the second visit which is likely to be an indication that by the second visit, all AD patients had progressed which would make the task of distinguishing them from the HC group easier. The lower accuracy for the third visit is caused by two subjects who were initially (visits 1 and 2) labelled as HC and MCI but who by the third visit had developed AD. So, they will be less severe than the other subjects. For the other scenarios (HC vs MCI and AD vs MCI), the classification accuracies also range at a similar level above 93% with the SVM outperforming SGD for most tasks.

Whereas the MMSE clinical scores prediction. for this particular task and based on the proposed regression model, the expectation is to have a better performance in the following visits. The results are presented in terms of the mean absolute error (MAE). The prediction within the visits, Figures 5.2, 5.3 and 5.4 shows the actual vs the predicted MMSE scores. The first three evaluation scenarios the MAE were 3.1, 2.6, and 3.7 for MMSE1, MMSE2 and

Table 5.4 Classification accuracies for three scenarios and visits original.

Scenario	Visit	Classifier	Accuracy %	F1_score	Precision	Recall	Confusion Matrex	
HC vs AD HC (n=42) AD (n=16)	1	SVM	86.2	0.86 (0.90/0.76)	0.86 (0.92/0.72)	0.86 (0.88/0.81)	**37** 3	5 **13**
		SGD	**91.3**	0.91 (0.94/0.83)	0.91 (0.93/0.86)	0.91 (0.95/0.81)	**40** 3	2 **13**
	2	SVM	**94.8**	0.94 (0.96/0.90)	0.94 (0.95/0.93)	0.94 (0.97/0.87)	**41** 2	1 **14**
		SGD	**94.8**	0.94 (0.96/0.90)	0.94 (0.95/0.93)	0.94 (0.97/0.87)	**41** 2	1 **14**
	3	SVM	84.4	0.84 (0.88/0.74)	0.85 (0.92/0.68)	0.84 (0.85/0.81)	**36** 3	6 **13**
		SGD	**86.2**	0.86 (0.90/0.77)	0.87 (0.94/0.70)	0.86 (0.85/0.87)	**36** 2	6 **14**
HC vs MCI HC (n=42) MCI (n=6)	1	SVM	91.6	0.90 (0.95/0.60)	0.90 (0.93/0.75)	0.91 (0.97/0.50)	**41** 3	1 **3**
		SGD	**93.7**	0.93 (0.96/0.72)	0.93 (0.95/0.80)	0.93 (0.95/0.66)	**41** 2	1 **4**
	2	SVM	**91.6**	0.91 (0.95/0.66)	0.91 (0.95/0.66)	0.91 (0.95/0.66)	**40** 2	2 **4**
		SGD	**91.6**	0.90 (0.95/0.6)	0.90 (0.93/0.75)	0.91 (0.97/0.5)	**41** 3	1 **3**
	3	SVM	91.6	0.92 (0.95/0.71)	0.93 (0.97/0.62)	0.91 (0.92/0.83)	**39** 1	3 **5**
		SGD	**95.8**	0.95 (0.97/0.83)	0.95 (0.97/0.83)	0.95 (0.97/0.83)	**41** 1	1 **5**
AD vs MCI AD (n=16) MCI (n=6)	1	SVM	**95.4**	0.95 (0.96/0.90)	0.95 (0.94/1.0)	0.95 (1.0/0.83)	**16** 1	0 **5**
		SGD	90.9	0.90 (0.94/0.80)	0.91 (0.88/1.0)	0.90 (1.0/0.66)	**16** 2	0 **4**
	2	SVM	**90.9**	0.90 (0.94/0.80)	0.91 (0.88/1.0)	0.90 (1.0/0.66)	**16** 2	0 **4**
		SGD	**90.9**	0.91 (0.93/0.85)	0.93 (1.0/0.75)	0.90 (0.87/1.0)	**14** 0	2 **6**
	3	SVM	95.4	0.91 (0.93/0.85)	0.93 (1.0/0.75)	0.90 (0.87/1.0)	**16** 1	0 **5**
		SGD	**96.4**	0.95 (0.96/0.92)	0.96 (1.0/0.85)	0.95 (0.93/1.0)	**15** 0	1 **6**

MMSE3, respectively. In visit 2, the error in prediction reduced compare to visit1. While MAE for visit 3 unexpectedly increased to 3.7 which again is likely to be caused by the two re-assigned subjects. In addition, Figures 5.5, 5.6 and 5.7 demonstrate the performance of the

Table 5.5 Classification accuracies for three scenarios and visits using SMOT (balanced groups) dataset.

Scenario	**Visit**		**Accuracy %**	**F1_score**	**Precision**	**Recall**	**Confusion Matrex**	
HC vs AD HC (n=42) AD(n=42)	1	SVM	**94**	0.94 (0.94/0.93)	0.94 (0.93/0.95)	0.94 (0.95/0.92)	**40** 3	2 **39**
		SGD	92.8	0.92 (0.92/0.92)	0.92 (0.92/0.92)	0.92 (0.92/0.92)	**39** 3	3 **39**

HC vs AD
HC(n=42)
AD(n=42)
	2	SVM	**97.6**	0.97
				(0.97/0.97)
					0.97
					(1.0/0.95)
						0.97
						(1.0/0.95)
							42
							2
								0
								40

		SGD	96.4	0.96 (0.96/0.96)	0.96 (0.95/0.97)	0.96 (0.97/0.95)	**41** 2	1 **40**

HC vs AD
HC(n=41)
AD(n=41)
	3	SVM	**91.4**	0.91
				(0.91/0.91)
					0.91
					(0.92/0.90)
						0.91
						(0.90/0.92)
							37
							3
								4
								38

		SGD	**91.4**	0.91 (0.91/0.91)	0.91 (0.92/0.90)	0.91 (0.90/0.92)	**37** 3	4 **38**

HC vs MCI
HC(n=42)
MCI(n=42)
	1	SVM	95.2	0.95
				(0.95/0.95)
					0.95
					(0.93/0.97)
						0.95
						(0.97/0.92)
							41
							3
								1
								39

		SGD	**96.4**	0.98 (0.98/0.98)	0.98 (0.97/1.0)	0.98 (1.0/0.97)	**42** 1	0 **41**

HC vs MCI
HC(n=42)
MCI(n=42)
	2	SVM	**98.8**	0.98
				(0.98/0.98)
					0.98
					(1.0/0.97)
						0.98
						(0.97/1.0)
							41
							0
								1
								42

		SGD	97.6	0.97 (0.97/0.97)	0.97 (0.97/0.97)	0.97 (0.97/0.97)	**41** 1	1 **41**

HC vs MCI
HC(n=41)
MCI(n=41)
	3	SVM	96.3	0.96
				(0.96/0.96)
					0.96
					(0.95/0.97)
						0.96
						(0.97/0.95)
							40
							2
								1
								39

		SGD	**97.5**	0.97 (0.97/0.97)	0.97 (0.97/0.97)	0.97 (0.97/0.97)	**40** 1	1 **40**

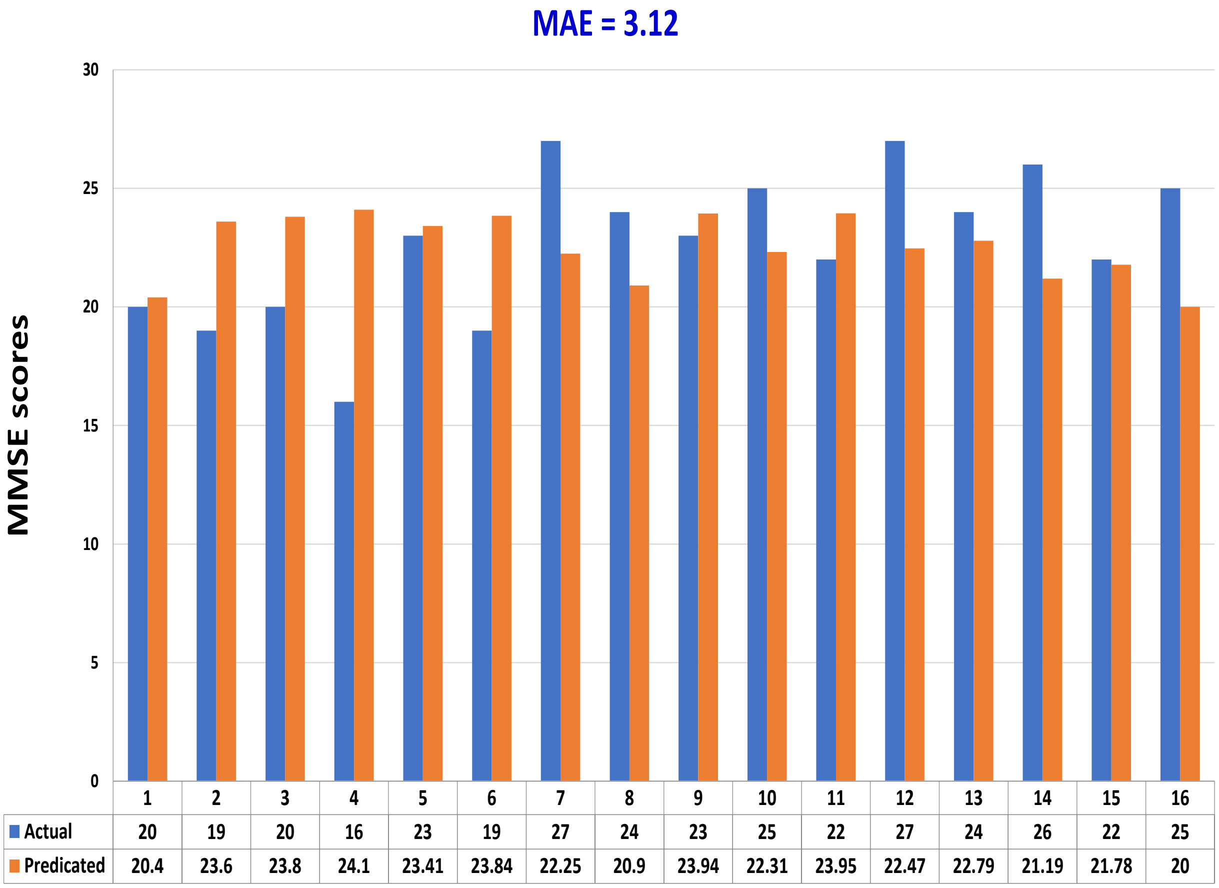
AD vs MCI
AD(n=16)
MCI(n=16)
	1	SVM	**96.8**	0.96
				(0.96/0.96)
					0.97
					(1.0/0.94)
						0.96
						(0.93/1.0)
							15
							0
								1
								16

		SGD	**96.8**	0.96 (0.96/0.96)	0.97 (1.0/0.94)	0.96 (0.93/1.0)	**15** 0	1 **16**

AD vs MCI
AD(n=16)
MCI(n=16)
	2	SVM	**93.7**	0.93
				(0.93/0.94)
					0.94
					(1.0/0.88)
						0.93
						(0.87/1.0)
							14
							0
								2
								16

		SGD	**93.7**	0.93 (0.93/0.93)	0.93 (0.93/0.93)	0.93 (0.93/0.93)	**15** 1	1 **15**
AD vs MCI AD (n=18) MCI (n=18)	3	SVM	**97.2**	0.97 (0.97/0.97)	0.97 (1.0/0.94)	0.97 (0.94/1.0)	**17** 0	1 **18**
		SGD	**97.2**	0.97 (0.97/0.97)	0.97 (1.0/0.94)	0.97 (0.94/1.0)	**17** 0	1 **18**

proposed cross-visits regression models, and showing the actual vs the predicted MMSE scores. The model that was built using visit1 samples was able to predict MMSE2 scores with a better



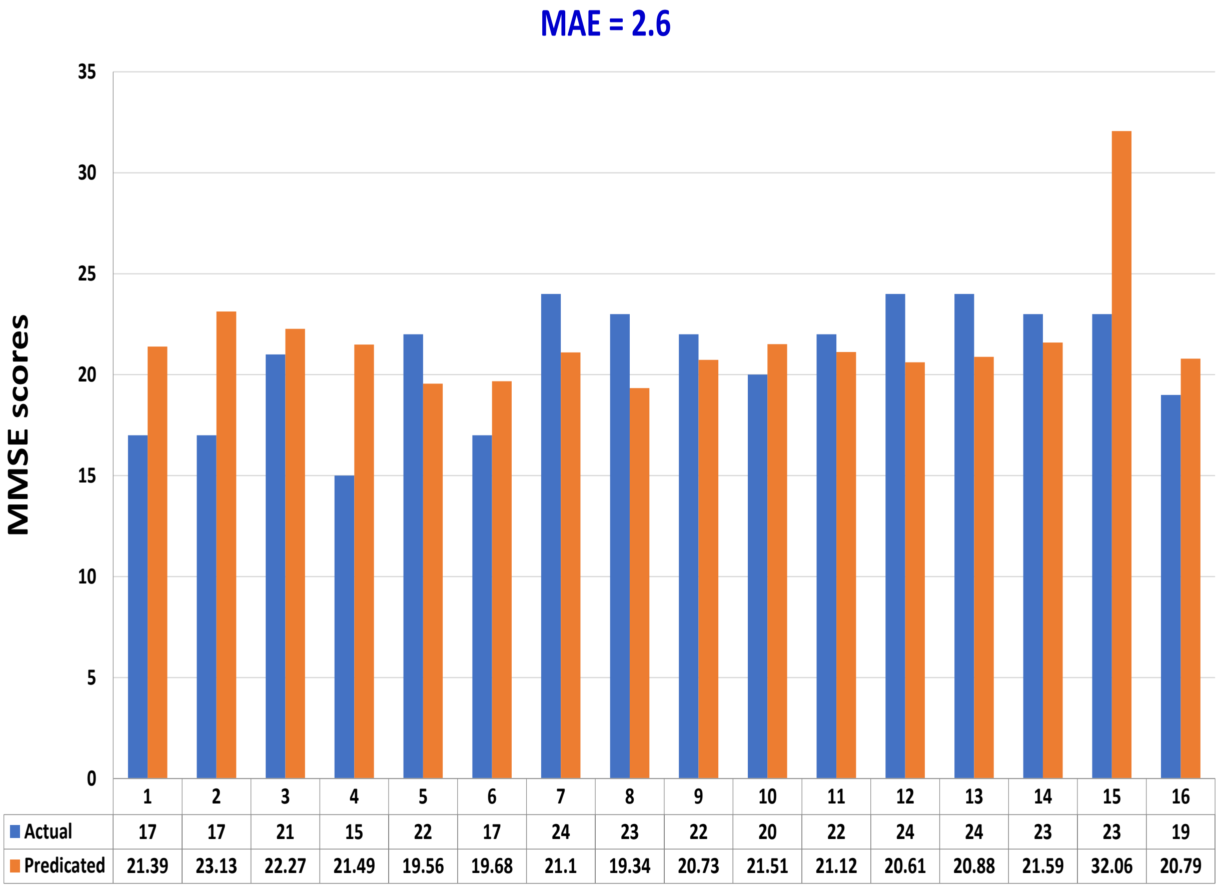
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
🟦 Actual	20	19	20	16	23	19	27	24	23	25	22	27	24	26	22	25
🟧 Predicated	20.4	23.6	23.8	24.1	23.41	23.84	22.25	20.9	23.94	22.31	23.95	22.47	22.79	21.19	21.78	20

Figure 5.2 Predicting MMSE1 - CV

MAE (1.18 compared to 2.6) for the visit 2 results. This might due to the close range of MMSE1 and MMSE 2 scores explains this improvement. Figure 5.6 and 5.7 also shows improvement in the performances at 2.25 and 2.18 compared to 3.7 (Fig 5.4) when estimating MMSE3 based on visit 2 data, and estimating MMSE3 using samples from both visit 1 and 2 respectively.

5.5 Discussion

This work aimed for longitudinal monitoring and predicting the disease severity; therefore, only 42 HC, 16 AD and 6 MCI participants were selected for those participants who returned for three visits. This selection created unbalance classes issue. Therefore SMOT technique was adopted to address this limitation. For AD class SMOT generated addational 26 synthetic samples to match the HC group, and similarly, for MCI it added 36 and ten samples to match HC and AD samples respectively. This technique not only address the performance in unbalanced

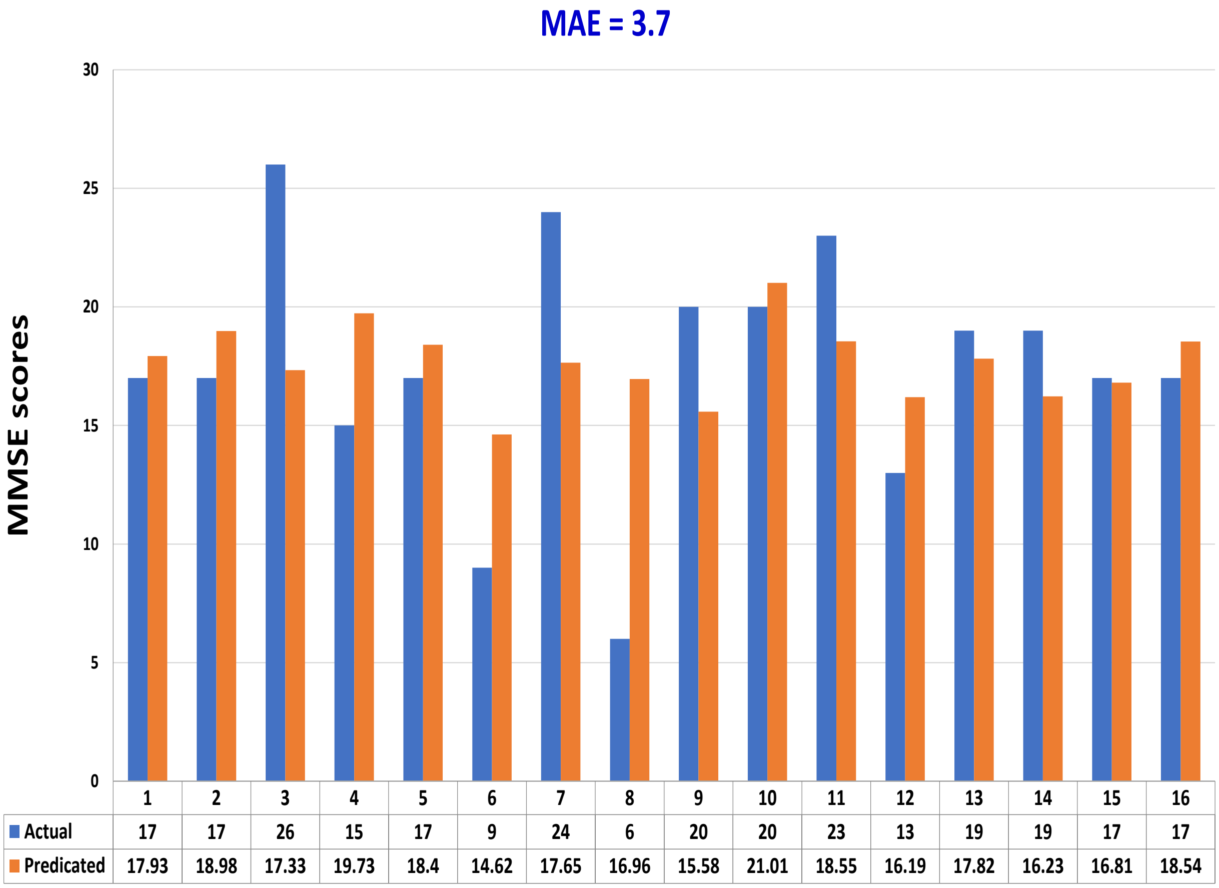


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
🟦 Actual	17	17	21	15	22	17	24	23	22	20	22	24	24	23	23	19
🟧 Predicated	21.39	23.13	22.27	21.49	19.56	19.68	21.1	19.34	20.73	21.51	21.12	20.61	20.88	21.59	32.06	20.79

Figure 5.3 Predicting MMSE2 - CV

classes but also improve the accuracies for all scenarios, and for the three visits, for example, classifying HC vs AD, with unbalanced configuration, the results were 91.3%, 94.8% and 86.2% compared to 94.0%, 97.6% and 91.4% for the first, second and third visits respectively.

Predicting MMSE clinical scores is performed for AD class only because the proposed model aimed to monitors AD condition, also the MCI group have few samples (only six participants). The expected performance in predicting MMSE scores for AD will be increased, i.e. MMSE prediction MAE error will be decreased longitudinally with visits, however, visit three unexpectedly was higher MAE of 3.7 compared to 3.12 and 2.6 for 1^{st} and 2^{nd} visits respectively, the unexpected behaviour may caused by the sharp drop in MMSE scores for some samples, for example, two patients scores were 23 and 24 at the 2^{nd} visit and decreased respectively to 6 and 13 for 3^{rd} visit. Moreover, the result from other scenarios that designed to estimate the next MMSE visit scores based on the current visits, the best model obtained MAE

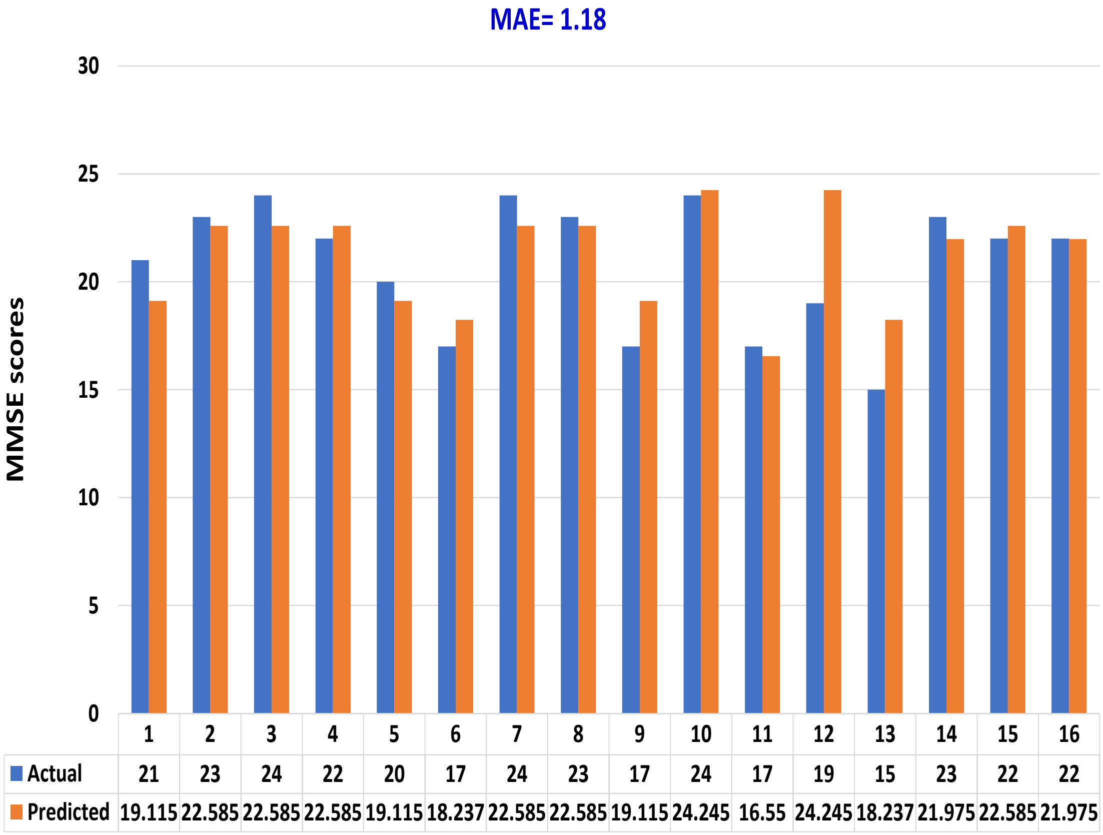


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
🟦 Actual	17	17	26	15	17	9	24	6	20	20	23	13	19	19	17	17
🟧 Predicated	17.93	18.98	17.33	19.73	18.4	14.62	17.65	16.96	15.58	21.01	18.55	16.19	17.82	16.23	16.81	18.54

Figure 5.4 Predicting MMSE3 - CV

of 1.18 when built using first visit's data to predict second visit MMSE. The improvement on estimating the future MMSE3 scores using the two scenarios (i.e., from visit2 and combined visit2 and 3 samples) may due to the fact that T was assumed to be equal to 365 days, which means the assigned expected MMSE score will be the same as the current value, for example, the sample with ID 075 Table 5.3, the estimated future MMSE3 score at 20 which is the same MMSE2 value. This is the case for 6 out of the 16 samples that missing visit3 examination date. Another attributable factor to this improvement was several samples have identical MMSE scores for subsequent visits, for example, samples 076, 091 etc. see Table 5.3 this make it an easy job for the classifier to correctly predict their MMSE scores.

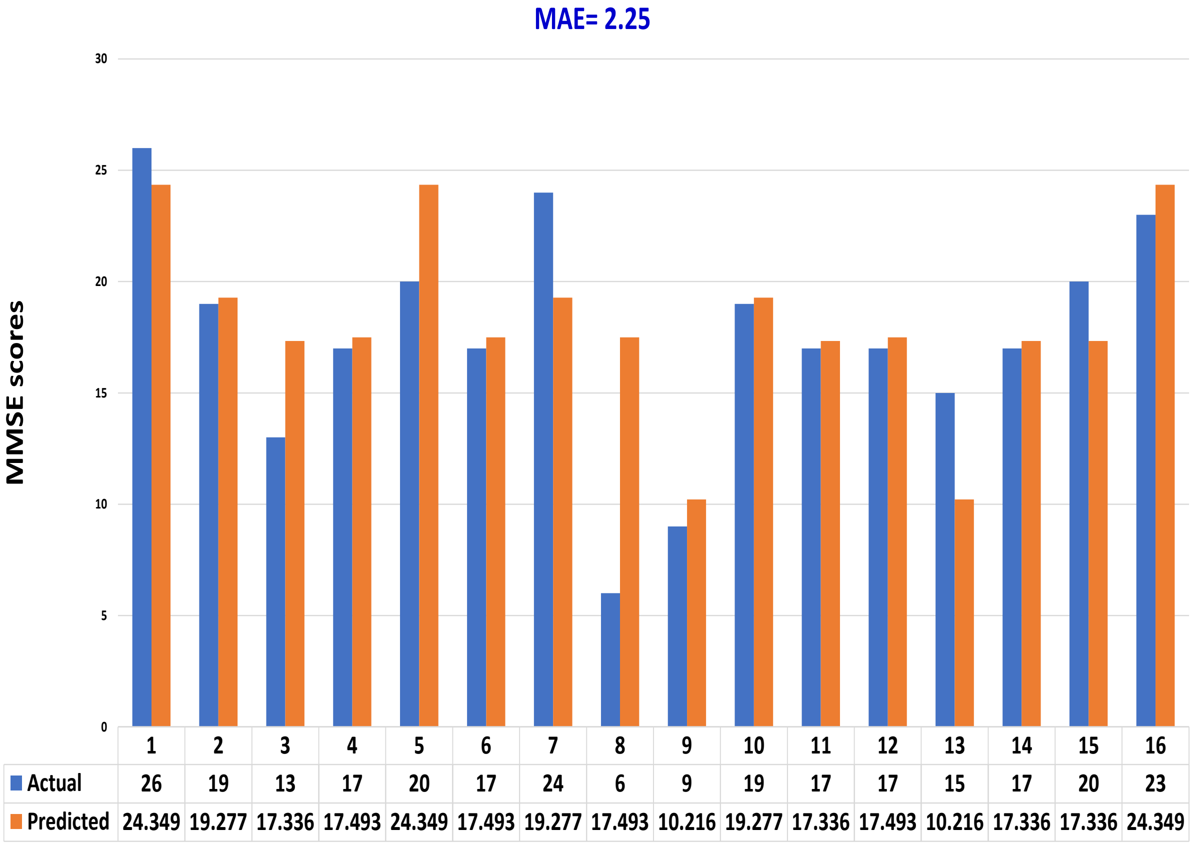
Comparing the proposed system to other studies from the literature. There were no studies performed a longitudinal classification task that used DementiaBank dataset. However, in terms of predicting MMSE, Yancheva *et al.* [170] conducted similar longitudinal AD severity inves-



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
🟦 Actual	21	23	24	22	20	17	24	23	17	24	17	19	15	23	22	22
🟧 Predicted	19.115	22.585	22.585	22.585	19.115	18.237	22.585	22.585	19.115	24.245	16.55	24.245	18.237	21.975	22.585	21.975

Figure 5.5 Predicting MMSE2 from Model build using MMSE1

tigation. They achieved MAE of 3.8 using the whole dataset, 2.9 for all samples attended more than three visits, 3.4 for all samples complete two visits, and 4.4 for all samples who did one visit. Yancheva *et al.* showed that longitudinal improvement on MMSE estimation despite that fact that some samples have increased in MMSE compared to the next visit, for example, sample 010 whos MMSE scores increased from 20 at visit1 to 21 at visit2 and later to 26 for visit3. Whereas the proposed system achieved MAE of 3.12, 2.6 and 3.7 for only 16 samples attended three visits, the system addresses the sharp drop in MMSE scores between the subsequent visits which explain the unexpected decreased in the performance for visit3. Furthermore, the proposed future MMSE estimation system further improved the performance with a limitation result from missing the examination date for some samples.

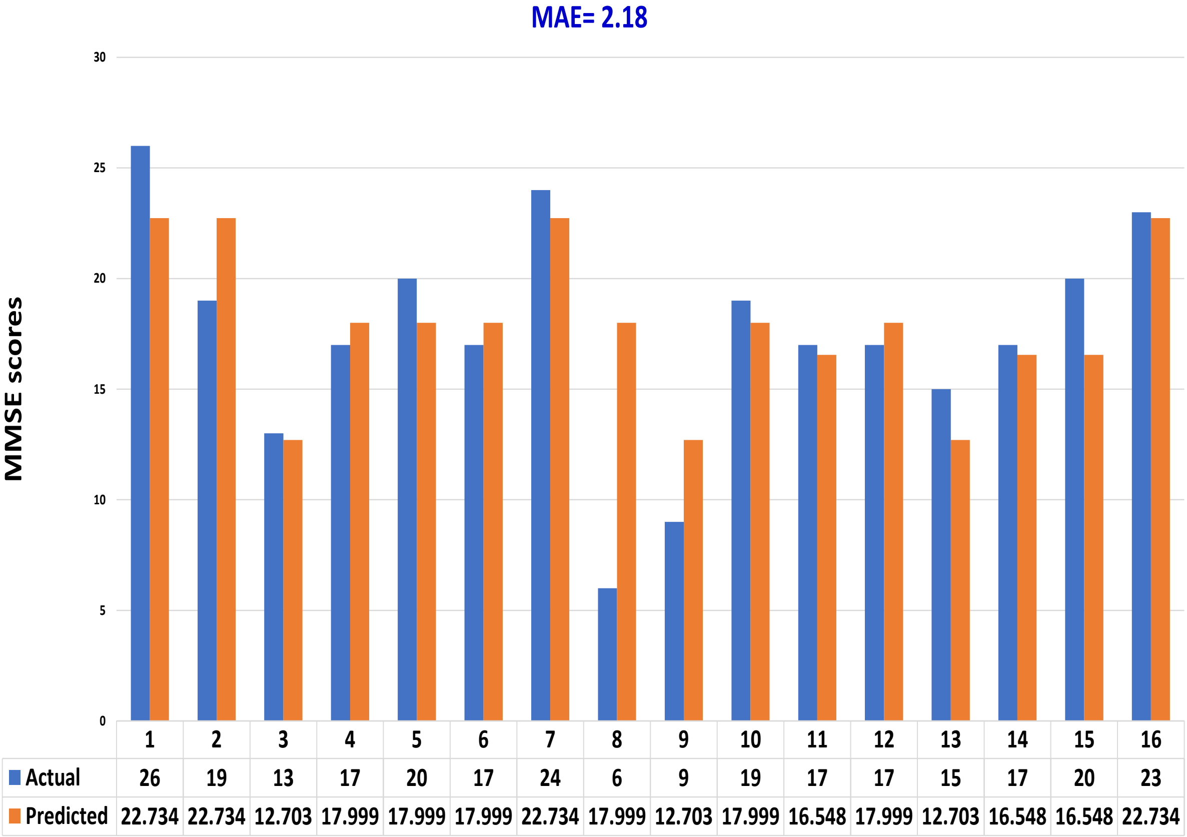


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
■ Actual	26	19	13	17	20	17	24	6	9	19	17	17	15	17	20	23
■ Predicted	24.349	19.277	17.336	17.493	24.349	17.493	19.277	17.493	10.216	19.277	17.336	17.493	10.216	17.336	17.336	24.349

Figure 5.6 Predicting MMSE3 from Model build using MMSE2

5.6 Summary

The proposed system achieved a promising results in estimating the MMSE scores. The results manifest the assumption of using speech features extracted from audio recordings for patients performing short vocal task to estimate the severity of AD. Several configurations were implemented to investigate the efficacy of the proposed system. These were designed based on longitudinal samples derived from the DementiaBank dataset. Predicting MMSE scores applied for only 16 out of 196 samples collected from patients who completed three visits. SMOT technique was introduced to address the unbalanced classes (AD n = 16, HC n = 42 and MCI n = 6) for the classification task, and SMOT also improved the classfication accuracy for all visits for all scenarios (AD vs HC, AD vs MCI and MCI vs HC). The nested LOOCV method was utilised instead of the k-folds due to the small sample size which in this case the recommended validation method [221]. The proposed extended spectral features were informative in both



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
■ Actual	26	19	13	17	20	17	24	6	9	19	17	17	15	17	20	23
■ Predicted	22.734	22.734	12.703	17.999	17.999	17.999	22.734	17.999	12.703	17.999	16.548	17.999	12.703	16.548	16.548	22.734

Figure 5.7 Predicting MMSE3 from Model build using MMSE1+MMSE2

tasks; predicting MMSE and discriminating between the three conditions.

In this chapter, two machine-learning models solely based on acoustic features were developed. The first model aims to distinguish between three groups of subjects representing three different cognitive levels: HC, MCI and people with AD condition. Using balanced datasets, the average cross visit SVM's accuracies were 94.3%, 96.7% and 95.9% when classifying HC vs AD, HC vs MCI and AD vs MCI respectively. The second model, presented new approaches for predicting the MMSE clinical scores, with average cross visit predictions MAE of 3.1. Furthermore, an investigated carried out to the contrasts in the results for the proposed MMSE prediction model. The acoustic features which used, were derived automatically from the speech recordings without the addition of any lexical or syntactic features that rely on complex speech recognition technology as in [170] . Finally, The finding and simplicity of the proposed models suggest that such methods can be used in the clinic and/or at home to guide the diagnosing and/or screening of dementia by using just speech.

Chapter 6

Detection early signs of neurodegenerative cognitive declines

6.1 Introduction

There has been a large increase in referrals from primary care of people with memory complaints to secondary care memory services, resulting in considerable pressure to diagnostic pathways [227]. The drive to seek early diagnostic clarification has led to an over 600% increase in referrals to secondary care memory clinics in the UK over the last ten years and generated considerable pressure on diagnostic pathways [227]. Although these dramatic changes have increased the number of patients in whom Neurodegenerative Disorders (ND) have been identified, a large proportion of the patients now referred to specialist memory clinics actually have Functional (non-progressive) Memory Disorder (FMD) concerns without objective evidence of cognitive deficits. Therefore, improvements to stratification and screening procedures would be highly desirable and could enable better targeting of limited health care resources [5].

When people visit a memory clinic, the assessment typically begins with a conversation with a specialist during which patients are asked a series of questions about their memory problems. This interaction provides important insights into the cognitive state of the patient. The clinician will note whether patient or accompanying others respond to questions, whether answers are quick and expansive or short and incomplete. Thus cheap, noninvasive and reliable stratification and screening tools which are fully automated, scalable, can be repeated and are remotely applicable are urgently required[5].

Recently, Mirheidari *et al.* [55, 228, 229] proposed an automatic method for the differentiation between patients with cognitive complaints due to ND or FMD inspired by diagnostic features initially described using the qualitative methodology of Conversation Analysis [60, 230]. In their work, a set of linguistic, acoustic and visual-conceptual features were extracted and used to train a number of classifiers. The highest classification accuracy of 97% was achieved using a linear support vector model. This work propose an automatic method to discriminate ND from FMD based solely on acoustic analysis of the same conversations used in [55]. The approach presented here differs from the approach pursued in the previous studies by Mirheidari et al. both in terms of complexity and in terms of the acoustic characteristics used. Here the system explore an acoustics-only approach based on data directly extracted from patients' speech signal. The prior study relied on more complex features (including features based on the contributions of clinicians and carers to the interaction) and required automatic speech recognition, natural language processing and natural language understanding.

6.2 Dataset

6.2.1 Participants

The dataset used in this experiment was recorded as part of a study conducted in the neurology-led memory clinic at the Royal Hallamshire Hospital in Sheffield, United Kingdom. Participants were recruited between October 2012 and October 2014 and the initial consultations between the neurologists and patients in the memory clinic were video and audio recorded. The study was approved by NRES Committee Yorkshire & The Humber - South Yorkshire. All participants were sent an information sheet prior to taking part in the study and given an opportunity to ask questions. They all gave written informed consent to participate and were informed that they could withdraw from the study at any time. Patients consented to their data being used for additional analyses by the research team but not to recordings of their interactions being made publicly available. All patients were referred because of memory complaints by General practitioner or other hospital consultant. Participants were encouraged to bring a companion such as carer or family member (if available) along to their appointment. Further details about the participant selection procedure have been provided previously[55]. At the clinic, patients

Table 6.1 Participants' details and test scores. ACE-R: Addenbrooke's Cognitive Examination-Revised; MMSE: Mini-mental state examination; PHQ9: Patient Health Questionnaire-9; GAD-7: Generalised Anxiety Assessment 7. Unpaired T-test was used. *ns** = not significant

	FMD(n=15)	ND(n=15)	Cut off	Max score	P-value
Age	57.8± 2.0	63.7 ± 2.3	N/A	N/A	p =0.06
Female	60%	53%			ns*
ACE-R	93.0 ± 1.4	58 ± 5.21	88	100	$p < 0.0001$
MMSE	28.9 ± 0.2	18.8 ± 2.0	26.3	30	$p < 0.0001$
PHQ9	5.6 ± 1.0	5.3 ± 2.0	5	27	ns
GAD7	4.7 ± 1.2	4.8 ± 1.5	5	21	ns
History taking part in minutes	range (10.1-32.3)	range(7.3-29.0)			

underwent a clinical assessment by a neurologist specialising in the diagnosis and treatment of memory disorders. In addition, all underwent the Addenbrooke's Cognitive Examination-Revised (ACE-R) cognitive assessment [231]. Neurologists also screened patients for clinical evidence of depression. Patients thought to be depressed clinically or patients with PHQ-9 [104] scores indicating a high risk of clinical depression were excluded from this study [55]. All participants also completed the Generalised Anxiety Disorder (GAD7) questionnaire [105] although high levels of anxiety were not considered an exclusion criterion. The final diagnoses of ND or FMD were formulated by Consultant Neurologists specialised in the treatment of cognitive disorders and also took account of brain Magnetic Resonance Imaging (MRI) findings and the result of a detailed separate neuropsychological assessment including the MMSE [61]; tests of abstract reasoning [232]; tests of attention and executive function [233]; category and letter fluency; naming by confrontation and language comprehension[234]; and tests of short and long-term memory (verbal and non-verbal) [235]. Table 6.1 gives an overview of participants' details and test scores. The ND group consisted of 10 cases of AD, 2 amnestic MCI, 2 BvFTD and one vascular dementia.

6.2.2 Diagnosis Process

Diagnoses of FMD were based on the criteria formulated by Schmidtke *et al.* [56] (although the proposed maximum age cut-off proposed by Schmidtke *et al.* was not applied). The participants in the ND group received the following neurological diagnoses: amnestic MCI as described by Petersen *et al.*[236], behavioral variant frontotemporal dementia as defined by Rascovsky *et al.* [237], and Alzheimer's disease diagnosis according to the NINCDS-ADRDA criteria [238].

Table 6.2 Details of the clinical session times expressed in minutes. STD: Standard deviation.

		Clinical session (Conversation +verbal fluency test)	**Conversation part only**	**Patient contribution to the conversation**
Mean time ± STD	FMD	34.3 ± 9.9	17.9± 8.5	11.5± 6.3
	ND	39.2 ± 8.0	19.4± 7.0	6.2± 4.5
Range time	FMD	(22.3 - 52.4)	(10.1 - 32.3)	(5.3 - 26.5)
	ND	(24.7 - 57.0)	(7.3 - 29.0)	(1.1 - 15.5)
Percentage	FMD	Not applicable	50.9 %	63.0 %
	ND	Not applicable	49.79 %	32.4 %

6.2.3 Memory clinic instructions

The neurologists, whose conversational activities elicited the dataset used in this study, followed a communication guide (developed on the basis of previously observed routine practice). Doctor-patient encounters began with a history-taking phase, which was followed by a brief cognitive examination (e.g. ACE-R), Table 6.2 provides timing details for the clinical sessions, history taking and percentage of the patients' contribution. Neurologists were encouraged to ask open questions to prompt conversation from the person with memory complaints. Examples of these questions includes the following: "When did your memory last let you down?", "Who is the most concerned about your memory – you or somebody else?", "Tell me a bit about yourself, where did you go to school?", "What did you do after you left school?", "Who looks after your finances?", "Do you smoke, have you ever smoked in the past?", "Why have you come to clinic today and what are your expectations?".

6.3 Proposed system

The system is intended as an early stratification tool for patients presenting with progressive ND-related cognitive problems based solely on diagnostic acoustic features in patients' speech. The proposed system pipline is similar to the system introduced in chapter 4 Fig 4.2, and it consists of three main stages: pre-processing, feature extraction and machine learning based classification.

6.3.1 Pre-processing

The clinical sessions were recorded using a "ZOOM H2N" portable digital recorder. The device was placed on the table between patient and doctor (within 1 m of neurologist, patient and accompanying person). The device produced audio files in “MP3” format with a sampling frequency of 44.1 kHz. The speech recordings were first converted to “wav” format and down sampled to 16kHz, and then inspected for background noise which may affect the quality of the extracted features. Similar denoising approach described earlier in chapter 4 section 4.3.1, was adopted and uses the Audacity(R) software [204] for both the audio conversion and denoising process. Before extracting any features from the recordings, the segments of the conversations containing the patient’s utterances (i.e. exclude those by the neurologist and any companions) were identified and isolated. For this work, the manually extracted turns were used as marked in the transcribed text files associated with the audio recordings.

6.3.2 Feature extraction

This work aimed to explore the potential of using only acoustic features to differentiate between FMD and ND. The authors of [55] applied a limited set of acoustic features inspired by the previous qualitative Conversation Analysis (mainly statistics of speech and silence). In the present study, a broader set of acoustic features were used. These features can be grouped into speech and silent features, phonation and voice quality features, and spectral features. Below, each main acoustic feature type is described in detail (see Table 6.3 for a summary of all features).

6.3.2.1 Speech and silent features

The frequency and duration of pauses have previously been reported to be of great value in the detection of dementia by means of speech analysis. In particular, it has been found that the speech of people with dementia is disrupted by more pauses compared to that of healthy people [54, 131, 144, 161, 210]. Based on this, ND patients expected to produce more and longer pauses, with shorter and fewer utterance periods compared to those with FMD. A silent segment of ≥ 0.25 seconds as a pause and a minimum voice segment of 0.5 seconds as a speech segment were considered. The Praat software [159] was used to identify pauses and speech segments in the recordings. A total of 14 features were used in the study including the max, mean and STD of the pauses and speech segments, the ratios of both max pause and speech segments and total time, the ratio of max speech and max pause segment. The last three features were the mean, STD, and variance of speech segments ≥ 0.8 seconds (i.e., excluding the filler words).

Table 6.3 Acoustic features.

Features	Type	Number of features
Speech and silent statistics	Speech and silent features	14
Fundamental frequency (F0) related measures	Phonation and voice quality	3
Harmonic-to-noise ratio (HNR)	Phonation and voice quality	3
Noise-to-harmonic ratio (NHR)	Phonation and voice quality	3
Shimmer$_{(local)}$	Phonation and voice quality	3
Jitter$_{(local)}$	Phonation and voice quality	3
Number of voice breaks	Phonation and voice quality	3
Degree of voice breaks	Phonation and voice quality	3
Average response time	Phonation and voice quality	1
Mel frequency cepstral coefficients (MFCC)	Spectral features	(extended to 5)
Filter bank energy coefficient (Fbank)	Spectral features	(extended to 5)
Spectral Subband Centroid (SSC)	Spectral features	(extended to 5)
Total		51

6.3.2.2 Phonation and voice quality features

This group of features is the same as described and extracted earlier in chapter 4; however, in this experiment, the mean, STD and variances were applied and used.

6.3.2.3 Spectral features

This group of features also were the same as described and extracted previously in chapter 4.

6.3.2.4 Statistical descriptive features

The dataset used in this study comprised of the part of routine outpatient encounters in which the doctor took the patient's history, the interactions lasted 18 minutes on average (see Table 6.2). The duration of the recordings allowed for the use of long-term features based on statistics calculated for this part of the conversation. Since the spectral coefficients are generated for each frame of all utterances, we first estimated their mean per utterance, and then we weighted them by dividing the averaged coefficients by the utterance time, and that produced the averaged weighted spectral coefficients (AWSC). The motivation behind that is the ND subjects are likely to provide fewer responses compared to FMD, so we believe that incorporating time factor will improve the predictive ability of the spectral features and hence increases the system accuracy. The statistical descriptive features are the mean, STD, min, max and the variance applied to each subgroup of the AWSC (i.e., MFFCs, Fbanks, and SSCs) resulting in adding an extra 15 features which makes the total number of features used in this experiment 51.

Table 6.4 Top (22) selected features using the wrapper, embedded and their statistical U-test. U: Mann-Whitney u-tests. Sample size n_1=n_2=15.

Rank	Features	U	P	Rank	Features	U	P
1	Mean time of all speech segments excluding filler words	17.0	0.00007	12	Mean response time	42.0	0.004
2	Ratio of max speech segment to the max pause time	19.0	0.0001	13	VAR degree of voice breaks	44.0	0.004
3	STD of total speech segments time excluding filler words	19.0	0.0001	14	STD of number of voice breaks	44.5	0.004
4	STD of the speech segments time	20.0	0.0001	15	Mean degree of voice breaks	45.5	0.005
5	VAR of total speech segments time excluding filler words	20.0	0.0001	16	Mean of Fbank coefficients	49.0	0.008
6	Ratio of max pause time to the total turn time	24.0	0.0002	17	Min of Fbank coefficients	49.5	0.009
7	Ratio of total pauses time to the total turn time	24.5	0.0002	18	VAR of SSC coefficients	52.5	0.01
8	Ratio of total speech segments time to the total turn time	26.0	0.0003	19	STD of SSC coefficients	59.5	0.02
9	VAR of number of voice breaks	26.0	0.0003	20	STD of MFCC coefficients	60.0	0.03
10	Ratio of total No. of pauses to the total turn time	27.0	0.0003	21	VAR of Fbank coefficients	60.5	0.03
11	Mean number of voice breaks	30.0	0.0006	22	Mean of MFCC coefficients	63.0	0.04

6.3.3 Feature selection

Feature selection (FS) as explained previously in chapters 4 and 5 is the process of selecting a subset of original features in order to optimally reduce the feature space according to a certain evaluation criterion [216, 217].

In general, there are three feature selection techniques namely the filter, wrapper and embedded [216, 217, 239]. The filter approach uses a specific ranking criterium (for example the Pearson correlation coefficient) to generate scores for each feature. The main advantages of the filter method are the low computational cost and speed compared to the wrapper approach; however, the feature ranking is done independently of the model's predictive ability, and that often leads to a loss of performance. The wrapper technique is computationally more expensive and slower compared to the filter approach. However, the wrapper usually results in improved performance because it utilises the model's ability to rank and select the best subset of features. Further, the wrapper method applies a learning algorithm known as the evaluator, and a search technique to find the combination achieving maximum model performance. The embedded method, on the other hand, executes the feature selection as part of the learning procedure, for instance, tree classifiers have a built-in feature importance identification capability and can therefore select the best subset of features.

A wrapper method was used based on an SVM evaluator known as the recursive feature elimination (RFE) technique [218], in which, the features are eliminated sequentially and the model performance estimated each time until all features have been excluded. The feature that has the maximum negative impact on the result is considered to be the most important one. Likewise, the rest of the features are then ranked. When using tree classifiers (random forest

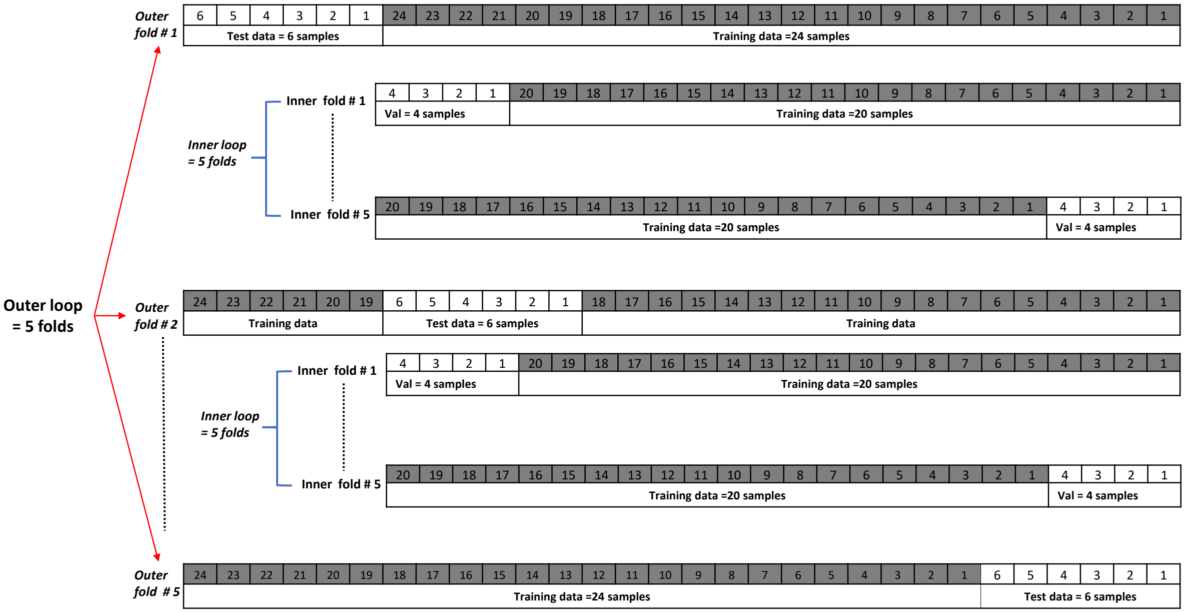


Figure 6.1 Nested k-fold cross validation with K=5.

and Adaboost) the built-in feature selection was utilised.

Another way of selecting features is by performing a statistical analysis, which measures the significance level of the mean difference between the FMD/ND classes. The null hypothesis assumes that there is no significant difference between the means for a particular feature and hence, that feature should be ignored. On the other hand, features that reject the null hypothesis are selected. For this task, the SPSS software [219] was utilised to perform Mann-Whitney u-tests appropriate for non-parametric data. Table 6.4 shows the best subset of features selected using both RFE and the embedded approaches as well as the corresponding p-values (below the 0.05 significance level) at a 95% confidence interval.

6.3.4 Validation Scheme

The same nested k-fold cross validation which explained earlier in chapter 4 section 4.3.4, however the number of fold k=5, Fig 6.1 shows the design of the nested 5-fold cross-validation. Feature selection and the model's hyper-parameter tuning were explored and the model with the best features and best parameters was tested using the test folds.

This work also explores another scenario, in which increase the sample size from 30 to 230 samples by partitioning each recording into 1-minute segment lengths, and extracted all features as mentioned in the previous sections. In this way the results were evaluated using a larger dataset, however, the balance between the two classes was lost due to the fact that the talk captured from conversations in the FMD group was longer compared to that from the ND group (i.e. 60% of the 230 samples were from the FMD group). The same principle of nested cross-validation was used as explained before in Fig 6.1, but changed from the k-fold method to a leave one group out cross validation (LOGOUT). The LOGOUT method guarantees that the group of samples that belong to one patient will always be in the same fold, for example, all segments from recording number 1 will be in the training data. The rational for selecting LOGOUT instead of the tradition LOO because the latter will not guarantee that the speech segments for one patients will be in one fold (i.e. a high probability that group of segments which belong to one patient will spread in all folds) and that will generate a bias in the model performance. LOGOUT also differed from the k-fold in generating the partitions. In LOGOUT the training data will be (G-1), and the test data will be the last remaining (G), (G) is the number of groups (30 in this case), this will loop again but now 30 times instead of five as in k-fold. The feature selection, model hyper-parameter tuning, the best model selection procedure, and the reported results remained the same as those calculated by the procedure outlined earlier. In this work the feature normalisation process was utilised using the following equation.

$$Stand(X) = \frac{x - \mu}{\sigma} \tag{6.1}$$

Where μ and σ is the mean and standard deviation of the training samples respectively.

6.4 Results

The results of the study suggest that machine learning models based on the analyses of the acoustic data from patients with cognitive complaints are capable of detecting differences between the two classes, ND and FMD, in keeping with prior research [144, 157, 158, 240, 241]. The discriminating potential of acoustic features was explored using five different classification algorithms (SVM, random forest, Adaboost, multi-layer perceptron, and SGD) and tested our findings using the validation procedure described above. The best models' results are listed in Tables 6.5 and 6.7 These were obtained using both scenarios: the original dataset with 30 samples and the augmented dataset with 230 samples.

Table 6.5 First scenario classification accuracies under different feature subsets and classifiers.

Classifier	**Accuracy using All features (51)**	**Accuracy with feature selection**	**No. of selected features**	**Accuracy using features with significance statistical P-value**	**No. of selected features**
Linear SVM	**93.0** ± 0.16 **%**	**97.0** ± 0.13 **%**	9	**97.0** ± 0.13 **%**	11
Random forest	90.0 ± 0.27 %	**97.0** ± 0.13 **%**	11	**97.0** ± 0.13 **%**	15
Adaboost	85.0 ± 0.40 %	93.0 ± 0.16 %	11	90.0 ± 0.24 %	21
MLP	**93.0** ± 0.16 **%**	**97.0** ± 0.13 **%**	20	**97.0** ± 0.13 **%**	22
Linear via SGD	90.0 ± 0.16 %	**97.0** ± 0.13 **%**	14	**97.0** ± 0.13 **%**	21
Mean	**90.2 %**	**96.2 %**	13	**95.6 %**	18

The average results for all models improved regardless of feature selection method and for both Tables 6.5 and 6.7. All models scored 97% accuracy except Adaboost which reached a maximum at 93% when the original dataset of 30 recording samples was used. The number of features used for each model is smaller when the wrapper and embedded approaches are used compared to the statistical ranking approach, for example the SVM wrapper model needed only 9 out of 20 ranked features from Table 6.4 compared to 11 features when the ranking is performed based on statistical significance. This differences between the two feature selection approaches were expected because both methods utilised the classifier scores to identify the best set of features maximising the performance while the analytical approach, on the other hand, included an un-optimised set of features for the models to reach their maximum accuracies. Table 6.7 shows the models' results when evaluated using the second scenario dataset. Both the SVM and SGD models score 92.0%, on average the five models achieved 90% which is less compared to 96.2% average results for the first scenario. The difference between the results was expected because the five models of the second scenario were evaluated on much more

Table 6.6 Extra classification metrics for the first scenario and for the same classifiers that used (wrapper and embedded) features selection approach.

Classifier	**Accuracy %**	**F1_score**	**Precision**	**Recall**	**Confusion Matrex**	
Linear SVM	**97.0**	0.966 (0.966/0.966)	0.968 (1.0/0.937)	0.966 (0.933/1.0)	14	1
					0	15
Random Forest	**97.0**	0.966 (0.966/0.966)	0.966 (0.933/1.0)	0.968 (1.0/0.937)	15	0
					1	14
Adaboost	93.0	0.933 (0.933/0.933)	0.933 (0.933/0.933)	0.933 (0.933/0.933)	14	1
					1	14
MLP	**97.0**	0.966 (0.966/0.966)	0.968 (1.0/0.937)	0.966 (0.933/1.0)	14	1
					0	15
Linear via SGD	**97.0**	0.966 (0.966/0.966)	0.968 (1.0/0.937)	0.966 (0.933/1.0)	14	1
					0	15

data samples, so the models' error percentage is likely to increase, however, this result may be considered more realistic and generalised thus perform better in future deployment.

6.5 Discussion

This study has shown that automatic speech analysis technology focusing and acoustic features in their speech could be a valuable complementary method in the diagnostic pathway of patients presenting with cognitive concerns. The work aimed to build a machine learning model that learns from our data and is able to predict the cognitive status of patients referred to a specialist memory clinic. In this study a binary classification was used; FMD or ND. The highest classification accuracy reached 97% which was achieved by four machine learning diagnostic models: SVM, random forest, multi-layer perceptron, and SGD. These features include ratios and statistics of pauses and utterances, which aligns with the literature. ND patients' speech has previously been found to be characterised by an increased number and duration of pauses as well as a reduction in the number of utterances which may be caused by difficulty in word finding (lexical retrieval)[210]. Similarly Singh *et al.* [53] and Roark *et al.* [54] reported that the mean time of both pauses and speech are useful in discriminating healthy subjects from MCI and AD patients. Other features of discriminating value in our study included the number and degree of voice breaks, which aligns with findings also previously reported by Meilán *et al.* [52].

Further, the importance of the spectral features including the average of Fbank and MFCC coefficients and the standard deviation of MFCC and SSC coefficients. These features measure the energy variations between frequency bands of a speech signal. As such they are harder to interpret in the context of detecting neurodegenerative cognitive decline. However, these features also capture some articulatory information expressed by lower resonance in the vocal tract, a prominent finding in ND patients (Fraser *et al.* [34] and in our previous works in [240, 241]). Some of the acoustic features we examined in this study were excluded from the discriminatory models finally created because they did not differe significantly between the two groups. These features include the fundamental frequency, shimmer, jitter and harmonic to noise ratio which appears to be in contrast to what have been stated in [52]. This may be because these features examine characteristics of the speech not strongly affected by FMD and ND, unlike in Parkinson's disease, where dysphonia is a common and key clinical feature [209, 222].

Comparing this model to that used in the previous study using the same dataset [55], the currently proposed method, based exclusively on acoustic findings, is computationally much less demanding than an analysis based on a combination of acoustic, lexical, semantic and visual-conceptual features. Furthermore, the previous classification approach included input features from the neurologist, and from accompanying persons whereas the present study only uses utterances from patients themselves. Although only patients' contributions and the analysis of acoustic signals were used in the present study, the overall classification accuracy improved to that achieved using the more complex approach (proposed 96.2% vs 95.0% [55])

The sensitivity and specificity of the proposed system were 93.75% and 100% respectively. This compares well with other dementia screening modalities, for example, electroencephalography (EEG) tests which may be relatively cheap, noninvasive and widely available, but are still rather cumbersome, and, more importantly, only have a sensitivity of 70% for the detection of early Alzheimer's disease [242]. Positron Emission Tomography (PET), although associated with much higher sensitivity and specificity (both at 86.0%) is invasive, requires the injection of a radioactive tracer via a peripheral cannula, means that patients have to be fasting for four hours before the test and is very costly [243]. Single photon Emission Computed Tomography (SPECT) is another diagnostic tool capable of demonstrating changes early in the course of neurodegenerative disorders with high sensitivity (86.0%) and specificity (96.0%), but is as cumbersome and costly as PET and also exposes patients to a high dose of radiation [59]

How does the proposed method compare to currently available approaches for screening at the interface between primary and specialist care? The test most commonly used world wide is the Mini Mental State Examination (MMSE). It takes and average of 10 minutes to administer. Although the MMSE has high sensitivity (87.3%) and specificity (89.2%) scores it is not sufficiently sensitive in the early stages of dementia. What is more, it is influenced by patients' level of education [62]. The ACE-R requires 12-20 minutes to administer and performs at a similar level (sensitivity 94.0% and specificity 89.0%), however, it does not provide feedback on why a particular diagnosis may have been made [66]. The clock drawing test (CDT), despite taking only 2-3minutes to complete and having impressive screening performance (sensitivity

Table 6.7 Classification accuracies for the second scenario (augmented dataset).

Classifier	**Accuracy using All features(51 ± STD)**	**Accuracy with feature selection ± STD**
Linear SVM	**87.0** ± 0.12 %	**92.0** ± 0.18 %
Random forest	84.0 ± 0.23 %	88.0 ± 0.28 %
Adaboost	81.0 ± 0.26 %	87.0 ± 0.29 %
MLP	**86.0** ± 0.26 %	91.0 ± 0.14 %
Linear via SGD	**87.0** ± 0.11 %	**92.0** ± 0.16 %
Mean	**85.0 %**	**90.0 %**

92.8% and specificity 93.5%), is not frequently used as it does not test memory as such, which especially limits its usefulness in AD. What is more, the scoring may be tricky due to having 8 different assessment settings. The test is also not suitable for people with illiterate patients who can't perform paper and pencil tests [244]

Importantly, all tools described above are the state of the art in dementia screening and currently utilised worldwide. In contrast, this approach has, so far, only been tested on a small dataset, so its performance should not be generalised unless been validated with very large dataset. Having said that, the study highlight the significance of an acoustic-only based approach as a promising low cost diagnostic aid in assessment pathways of patients presenting with cognitive problems. In view of the relatively low hardware (microphone) and computational complexity this system could easily be deployed in settings other than memory clinics. This may be desirable from a patients' perspective and more cost-effective from the perspective of health care providers. Moreover, whereas the application of a system based on semantic understanding is limited to the language(s) it was trained to interpret, the present system, using acoustic features only, is much less dependent on the particular language used by a patient and should therefore be more widely generalisable.

There are several limitations to this study. Firstly the data set is relatively small with only 15 cases in each group. However, larger dataset model evaluated, and the resulting difference was only a 5% decrease in performance. However, the size difference between the two scenarios was more > 750% (30 samples compared to 230) i.e. for each 100% increment in size the performance decreased by 0.67% which shows that the proposed approach did not deviate badly. Also the high accuracy of this model reflects the difference between the two groups; notably the

ND group were mostly in the moderate stages of ND, with a mean MMSE of 18.8(+/−2.0). Furthermore, this approach was based on manually annotated files that identified the patients' turns in standardised conversations, however, this limitation can be overcome by using an automatic speaker diarization software which is capable of providing information about who is speaking and when. Further work is required including evaluating our model's performance with spontaneous conversations and with the use of speaker diarization software to fully automate the proposed system.

6.6 Summary

The results of this study lead to several conclusions. First, a relatively small number of extracted acoustical features are shown to be of great importance in the differentiation between ND and FMD. These features are likely related to changes in the neurobiology associated with a given neurodegenerative cognitive disorder, reflected in the acoustic output. Secondly, the proposed approach can be easily deployed at clinics and during standard clinical encounters. This will require only minimal effort on the part of the examiner and mean a much quicker diagnosis for the examinee. Finally, despite the limitations of this study, the findings show that acoustic-only features offer a potential low-cost, simple and alternative to more complex features requiring automatic speech recognition, part-of-speech parsing, and understanding of speech in the automated screening or stratification of patients with cognitive complaints. Hence it has great potential for use early in pathways that assess people with cognitive complaints.

Chapter 7

Depression assessment using acoustic features

7.1 Introduction

This chapter demonstrate the ability of newly developed acoustic features to detect person's level of depression. The proposed features will be used to build two machine learning systems. This work also show how these systems can be improved when fusing the new features with the state of the art acoustic features, which was previously used in detecting dementia (listed in Table 6.1). Typically Gradient boosting classifier (GBC) or Gradient boosting regressor (GBR) is then trained to operate on the acoustic feature space. The testing undertaken in this chapter aims to gain insights into the usefulness of the acoustic ability for predicting depression score.

Several challenges arise when building systems capable of classifying or predicting depression level in speaker's recording. One challenge is the lack to samples for each speaker at different levels of depression. Another problem is the diversity of recordings length (ranges from several seconds to 27 minutes), and the speech tasks (spontaneous and read, vocal exercises, and the recordings do not always contain all the tasks). Hence the expected phonetic variability inside and between files is large. This variation potentially affects the classification and prediction models and makes it a challenge to achieve satisfactory results. Therefore the new features derived based on the temporal speech characteristics will be investigated to eliminate such limitations. These features aim to capture utterance behaviour and thus under depression symptoms

Table 7.1 Depression datasets details

#	Dataset	Language	Total subjects	No. of samples	Task	Evaluation method	Data partitions	No. of subjects	Class 0 No. samples	Class 1 No. samples	Mean sessions in seconds
1	**AVEC-2013**	German	129	150	Multi	BDI	Training	42	26	24	845.68
						Class 0 ≤ 13	Development	43	26	24	845.24
						Class 1 otherwise	Testing	44	25	25	842.82
2	**AVEC-2014**	German	127	300	Q/A and reading	BDI	Training	42	52	48	52.6
						Class 0 ≤ 13	Development	43	52	48	45.12
						Class 1 otherwise	Testing	42	50	50	56.8
3	**AVEC-2016**	English	186	186	Conversation	PHQ	Training	107	77	30	433.43
						Class 0 <10	Development	32	22	10	490.83
						Class 1 otherwise	Testing	47	33	14	511.72

this behaviour could be identified, furthermore, incorporating the weighting scheme (explained in chapter 6 section 6.3.2.4) while extracting the acoustic features will minimise the effect of this unwanted variation. The analysis in this chapter will include a statistical investigation into the acoustic space, feature significance exploration and results comparison.

7.1.1 Datasets

The experimental results in this chapter were based on using three commonly published depression datasets. The first dataset is known as the Audio/Visual Emotion Challenge and Workshop (AVEC) 2013, herein called AVEC-2013. The following data is the Audio/Visual Emotion Challenge for the year 2014 noted as AVEC-2014, and the third dataset is the DAIC-WOZ corpus. These datasets are the only publicly accessible depression speech corpus; also, the challenge proposed by the owners is to pursuit an optimum solution for building objective evaluation models. The available benchmark results allow the researchers to compare their results with the baseline. The following sections will provide details for the datasets used. Also Table 7.1 provides general information regarding the datasets' partitions.

7.1.1.1 AVEC-2013

The AVEC-2013 depression data challenge was designed to seek competition among researchers to build robust machine learning depression-assessment models by utilising the video and audio recordings. The baseline results were provided in the form of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for both video and audio modalities see Table 7.7 [245]. The dataset encompass a total of 150 video and audio recordings and were available in the form of three partitions training, development and testing, and each part contains a total of 50 recordings. The data collected using headset connected to the built-in sound-card of a laptop at 44.1 KHz sampling frequency with 16 bit encoding. The recording length ranges between

50 and 20 minutes (mean = 25 minutes), and the sessions are recorded in a number of quiet settings. All participants in AVEC-2013 are native German speakers, with an age range of 18 to 63 years(mean 31.5 years, and a standard deviation of 12.3 years) [245]. The participants were instructed to accomplish several tasks such as vocal exercises (counting 1 to 10), spontaneous free form (FF) speech (e.g. talking loud during task solving, and narrating a story from their past), and read tasks (NW) (e.g. Northwind Passages from the novel Homo Faber by Max Frisch) [246]. The depression scores are obtained using the BDI depression inventory, and the average BDI score was 15.1 for the training partition, 14.8 for the development, and 14.5 for the test partition (standard deviations = 12.3, 11.8,and 11.4 respectively). BDI were provided as a single value for each file in the training, development and test partitions. Following the common practice stated in the BDI regulations [86], this work adopts a depressive symptomatology with a cut-off of 13/14 points on the BDI scale and this cut-off a fairly balanced distribution of speakers with (n = 96) (labelled as class 1, BDI $>$ 13) and without major depressive symptomatology (n = 104) (labelled as class 0, BDI $\leq$ 13).

7.1.1.2 AVEC-2014

This data is a subset of the previous AVEC-2013 challenge corpus, and it has a total of 150 videos of task-based depression data. The tasks were recorded in the form of virtual agent-human interactions and using a close-talk microphone and a webcam. Each recording contains a task performed by one speaker. A total of 84 participants were included in the dataset, and in some cases, participants had provided more than one sample [246]. The recording sessions were carried out in a rate between one and four times with two weeks break between the sessions. The sessions last between 50 and 20 minutes, with mean = 25minutes while the total recording time for the whole data was 240 hours. The mean age of the participants was 31.5 and ranged between 18 to 63 years, with a standard deviation of 12.3 years. The sessions recorded in quiet environments. Only two out of 14 recorded tasks were provided and for each part of the three partitions (train, develop and test) contains 100 recordings, and these recordings were 50/50 divided into two tasks. The sessions last between (6 seconds to 4 minutes and 8 seconds)[246]. The two selected tasks were as follows:

- Northwind - Participants, read loudly a passage of the story “Die Sonne und der Wind” (The North Wind and the Sun), expressed in the German language. Here and after this task will be referred as "NW".

• Freeform - Participants answer to one of several questions such as: “What was your best gift, and why?”, “What is your favourite dish?”, “Discuss a sad childhood memory”, and same in the German language. Here and after will be refer to this task as "FF".

Similar to AVEC-2013, the depression labels for the dataset were provided using BDI evaluation tool and the same dpression levels cut-offs were adopted, however, the baseline challenge performances listed in Table

7.1.1.3 AVEC-2016-DAIC-WOZ

This dataset includes clinical interactions constructed to assess the diagnostic of physiological distress disorders such as anxiety, depression, and post-traumatic stress disorder (PTSD). These interviews were obtained to build a computer agent that examines people and distinguishes verbal and nonverbal signs of mental illness [247]. The data collected combine video and audio recordings, and comprehensive questionnaire responses; this part of the corpus comprises the Wizard-of-Oz interviews, administered by an animated virtual interviewer named "Ellie", managed by a human interviewer in a different room. The DAIC-WOZ data was recorded by a close-talking microphone, Kinect, and camera, data already annotated and transcribed to help extract different types of features, such as verbal and non-verbal [247]. A total of 189 English native speakers participated in this study providing 187 speech recordings (some sessions discarded due to technical errors) and the recording sessions were between 5-20 minutes long, while the actual conversations time as explored in the experiment ranges from 1.02 to 21.22 minutes. The conversations were semi-structured, starting with standard questions to help establish a friendly environment during the session; later, the questions were directed specifically towards symptoms and issues linked to anxiety, depression, and PTSD. Bellow is a discourse sample between the virtual agent "Ellie" and a participant [247]:

Wizard-of-Oz conversation sample:

- Ellie: "Who’s someone that’s been a positive influence in your life?"
- Participant: "Uh my father".
- Ellie: "Can you tell me about that?"
- Participant: "Yeah, he is a uh."
- Participant: "He’s a very he’s a man of few words."
- Participant: "And uh he’s very calm."

- Participant: "Slow to anger."
- Participant: "And um very warm very loving man."
- Participant: "Responsible."
- Participant: "And uh he's a gentleman has a great sense of style and he's a great cook".
- Ellie: "Uh huh."
- Ellie: "What are you most proud of in your life?"

The data labelled with a depression level score for each recording, and it is obtained using the standard self-evaluation Patient Health Questionnaire (PHQ). Binary classes also given: 1 for depressed (PHQ $\geq$ 10) and 0 normal or no depression (PHQ $<$ 10). Variety of features also provided including video features such as facial landmarks, eyes, and head pose gaze directions, position, and orientation of the head, and a histogram of oriented gradients. While audio features include spectral (MFCC, and phase distortion), prosodic(F0, and voicing), and other voice quality measures.

7.2 Proposed system models

The general pipline for the proposed models is illustrated in Fig 7.1, both classification and regression models have the same backbone which starts with pre-processing and feature extraction steps, also they share the same training methodology and evaluation scenarios which will be explained in the following sections.

7.2.1 Feature pre-processing and extraction

Wide range of acoustic features proven to have a discriminating ability when utilised as diagnosis tool in variety of conditions including depression [99, 171, 174–177, 180, 188, 194], Parkinson's disease [207, 209, 211, 222], dementia [139, 144, 156, 160, 161, 166, 170, 208], autism [248–251], etc., and the previous works demonstrated that spectral, temporal and voice quality features were useful in diagnosing AD [240, 241] and discriminating FMD from ND conditions [252] thus the features which previously extracted in chapters 4, 5 and 6 (listed in Table 4.2) and combined with the newly developed features will be used in this experiment. Furthermore, additional set of spectral features will also be used and these include:

* Spectral roll-off (SRO): is the frequency following to which a specific cutoff ratio of the magnitude allocation of the spectrum is allocated. The equation for SRO is given by:

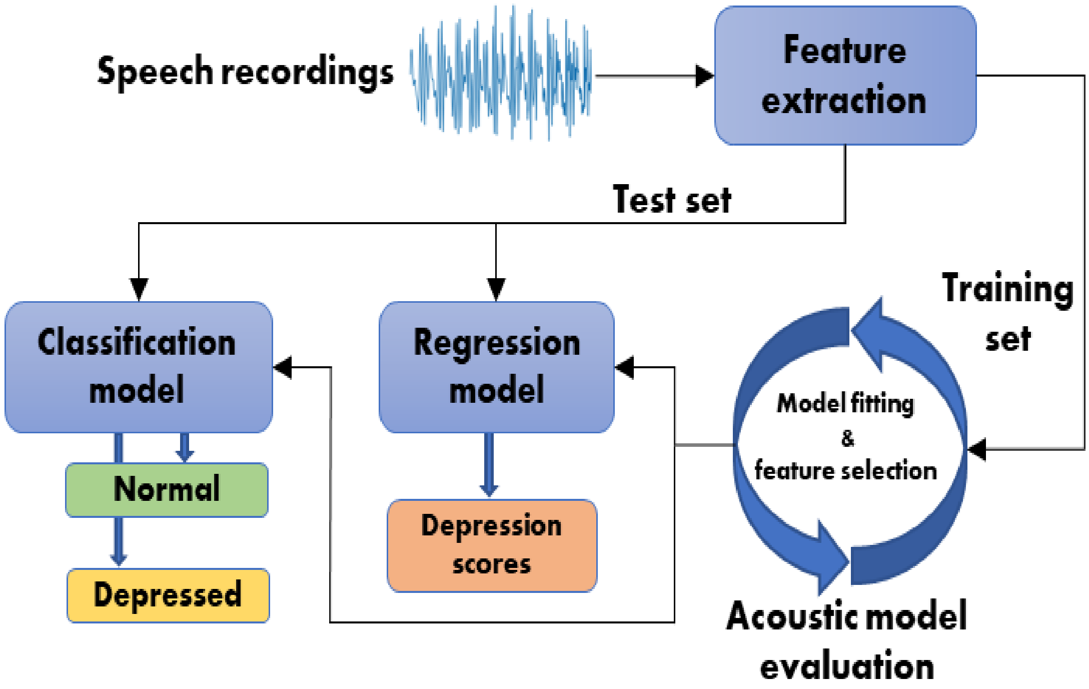


Figure 7.1 Proposed system for depression assessment

$$SRO = \frac{z}{100}\sum_{k=0}^{K-1}|S_i(k)| \tag{7.1}$$

where z is ranges between (80-90%), and in this work z is set to 85% (the default value [253]).

* Spectral entropy (SE): It measures the entropy for subframes' normalised spectral energies [253] computed as follows: first, is to normalise the power spectral density $P_i(k)$ so it can be viewed as probability density function given by:

$$p_i = \frac{P_i(k)}{\sum_0^i P_i(k)} \tag{7.2}$$

Then, the SE is estimated using the following formula for an entropy:

$$SE = -\sum_{i=1}^{n} p_i \ln p_i \tag{7.3}$$

* Spectral flux (SF): estimate how the spectral vary between two consecutive frames and

is estimated as the squared difference between the normalised magnitudes of the spectra of the two successive short-term windows [254], estimated using :

$$SF_{i,i-1} = \sum_{k=1}^{N} (\ p_i(k) - p_{i-1}(k)\ )^2 \tag{7.4}$$

* Spectral spread (SS): Is a measurement of average deviation of the spectrum from the spectral centroid [253] using the following equation:

$$SS = \sqrt{\frac{\sum_{k=a1}^{a2} (f(k) - SSC_i)^2 * S(k)}{\sum_{k=a_1}^{a_2} S(k)}} \tag{7.5}$$

where $f(k)$ represent the frequency in Hz equivalent to bin k, $S(k)$ is the spectral value at bin k, a_1 and a_2 are the band boundaries, in bins, over which to estimate the spectral spread. SSC is the spectral centroid.

7.2.1.1 Formants features

Formants are the most significant elements in the speech spectrum and accommodate substantial amounts of information on the resonance characteristics of the vocal tract. In the source-filter model of speech production, the vocal tract is represented as a time-invariant all-pole linear filter whose poles somewhat equivalent to vocal tract formants. The linear predictive analysis of this filter is the most used method for computing formants and their bandwidths [255]. In this experiment, the formants contour were extrated using Praat Boresma *et al.* [159] software, and included only the first two formants $F1$ and $F2$ and their bandwidths $B1$, $B2$ and intensity $I1$ and $I2$ respectively. The selection of these features was inspired by several studies [171, 178, 179], who reported the effectiveness of these features in the task of detecting depression condition.

7.2.1.2 New speech activity behaviour features

The newly developed features were motivated by the fact that temporal characteristics are influenced under certain conditions; for example, in depression condition, the patient reported producing shorter utterances and longer pauses [74, 128, 129]. However, these features frequently explored using the standard functions such as the mean, maximum, STD, skewness, kurtosis etc.[180], and sometimes it is difficult with just these features to reliably conclude a disease-related behaviour especially when the patients and healthy subjects share a close range

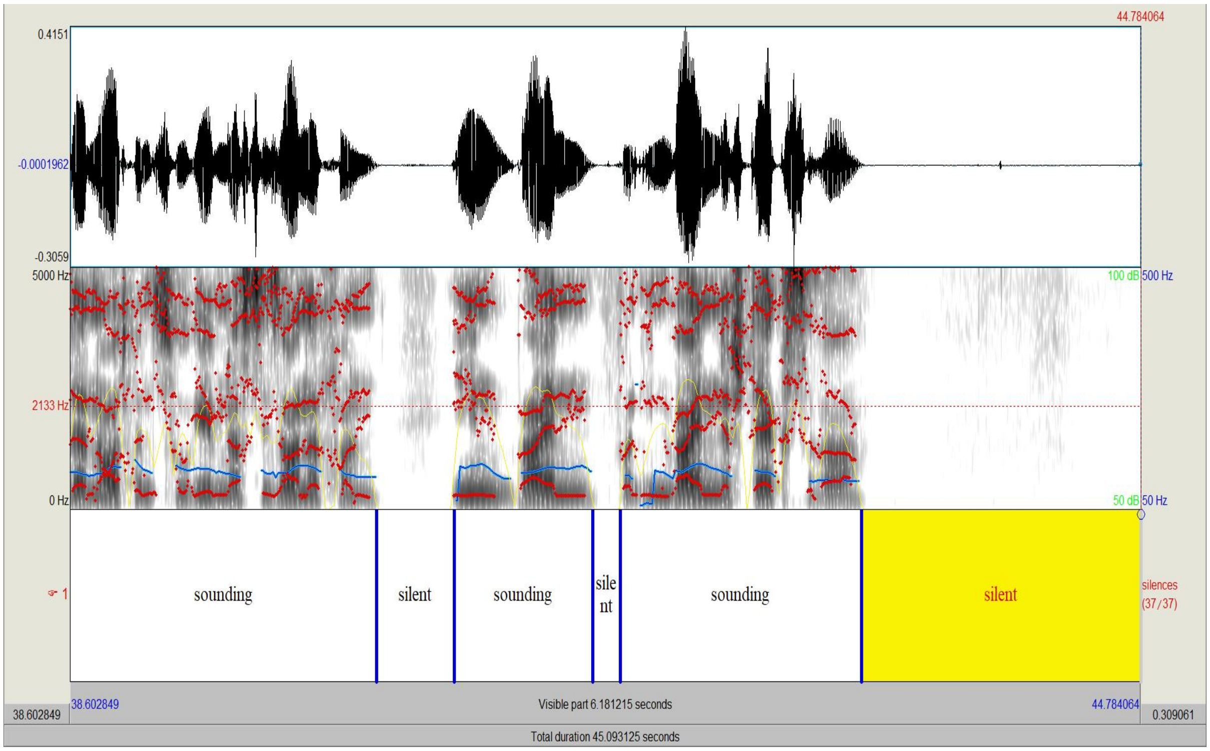


Figure 7.2 Praat text grid for voice and silent segments

of symptoms or conditions. Therefore it is highly desirable to find variables that can be used to infer behaviour associated with specific disease such as depression.

The new features extracted based on the average of both voiced and silent segments and as follows:

First by using Praat text grid silences, all voice and silent segments time (in msec) were identified in the recordings as illustrated in Fig 7.2. The result are two vectors, the first one contains voice segments time $\vec{V}$ =[v_1,v_2,...,v_N] for all v segments ≥ 0.5 msec, and the second vector is for pauses segments time $\vec{P}$ =[p_1,p_2,...,p_M] for all p pause segments ≥ 0.25 msec, where N, M are the total number of voice and pause segments respectively. Then from these vectors, several variables were computed:

The average of voice segments times estimated:

$$Mean\_V = \frac{\sum_{i=1}^{N} V_i}{N} \tag{7.6}$$

Maximum voice segments time:

$$Max\_V = Max\ \vec{V} \tag{7.7}$$

The standard deviation of voice segments time:

$$STD\_V = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(V_i - Mean\_V)^2} \tag{7.8}$$

The variance of voice segments time:

$$VAR\_V = \frac{1}{N}\sum_{i=1}^{N}(V_i - Mean\_V)^2 \tag{7.9}$$

The total voice segments time:

$$T_{time}\_V = \sum_{i=1}^{N} V_i \tag{7.10}$$

The total number of voice segments:

$$No\_V = N \tag{7.11}$$

The ratio of maximum voice time to the total voice segments time:

$$V_{max}R = \frac{Max\_V}{T_{time}\_V} \tag{7.12}$$

The average of pause segments times:

$$Mean\_P = \frac{\sum_{i=1}^{M} P_i}{M} \tag{7.13}$$

Maximum pause segments time:

$$Max\_P = Max\ \vec{P} \tag{7.14}$$

The standard deviation of pause segments time:

$$STD\_P = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(P_i - Mean\_P)^2} \tag{7.15}$$

The variance of pause segments time:

$$VAR\_P = \frac{1}{M}\sum_{i=1}^{M}(P_i - Mean\_P)^2 \tag{7.16}$$

The total voice segments time:

$$T_{time}\_P = \sum_{i=1}^{M} P_i \tag{7.17}$$

The total number of pause segments:

$$No\_P = M \tag{7.18}$$

The ratio of maximum pause time to the total pause segments time:

$$P_{max}R = \frac{Max\_P}{T_{time}\_P} \tag{7.19}$$

The total voice and pause segments time:

$$T_{time}\_VP = T_{time}\_V + T_{time}\_P \tag{7.20}$$

$$T_{No}\_VP = No\_V + No\_P \tag{7.21}$$

$$Mean\_VP = Mean\_V + Mean\_P \tag{7.22}$$

$$P_{max}R_T(VP) = \frac{Max\_P}{T_{time}\_VP} \tag{7.23}$$

$$V_{max}R_T(VP) = \frac{Max\_V}{T_{time}\_VP} \tag{7.24}$$

From the two vectors $\vec{V}$, $\vec{P}$ six new vectors were created, and five functions (mean, STD, VAR, total time and the total number of segments) were applied to the new vectors (estimated similarly using previous equations). Thus the new vectors and their new features expressed as follows:

1. Active voice vector $\vec{AV}$ =[v_1,v_2,...,v_I] for all voice segments V > Mean_V

   $\vec{AV}$ features = "$Mean\_AV$", "STD_AV", "VAR_AV", "T_{time}_AV" and "$No\_AV = I$"

2. Passive pause vector $\vec{PP}$ =[P_1,P_2,...,P_J] for all pause segments P > Mean_P

   $\vec{PP}$ features = "$Mean\_PP$", "STD_PP", "VAR_PP","T_{time}_PP" and "$No\_PP = J$"

3. Passive voice vector $\vec{PV}$ =[v_1,v_2,...,v_K] for all voice segments V < Mean_V

   $\vec{PV}$ features = "$Mean\_PV$", "STD_PV", "VAR_PV","T_{time}_PV" and "$No\_PV = K$"

4. Active pause vector $\vec{AP}$ =[P_1,P_2,...,P_L] for all pause segments P < Mean_P

   $\vec{AP}$ features = "$Mean\_AP$", "STD_AP", "VAR_AP","T_{time}_AP" and "$No\_AP = L$"

5. Cross voice vector $\vec{CV}$ =[v_1,v_2,...,v_R] for all voice segments V > Mean_P

   $\vec{CV}$ features = "$Mean\_CV$", "STD_CV", "VAR_CV", "T_{time}_CV" and "$No\_CV = R$"

6. Cross pause vector $\vec{CP}$ =[P_1,P_2,...,P_S] for all pause segments P > Mean_V

   $\vec{CP}$ features = "$Mean\_CP$", "STD_CP", "VAR_CP", "T_{time}_CP" and "$No\_CP = S$"

This work hypothesise that the new vectors and their features will be superior to the common functions which were used in the literature to describe voice and pauses such as the mean of pauses. Because these common features may not characterise specific behaviour for example in case of a small number of voice segments but larger in time may have similar average to a larger number of voice segments but shorter in time. Therefore the new features will provide deep insight into the aspects of these segments, which might make them more informative in terms of discriminating between healthy and depressed people due to the facts mentioned by Emil Kraepalin [128] that depressed people speak with longer pauses and hesitations, slowly and with low volume, and sometimes whispering. The new features will provide details of how active the person is when speaking, for example, this system expect that with "$Mean\_AV$" for normal subjects will have higher value compared to people suffer from depression. "$Mean\_AV$" measures the average of all voice segments produced by subjects and were larger than his mean of utterance i.e. "$Mean\_V$". Whereas "T_{time}_AV" measure how much amount of time that someone had spoken above his "$Mean\_V$".

In contrast to $\vec{AV}$ the $\vec{PP}$ the higher values of these vector features will infer that someone is speaking with larger pauses. Thus patients with depression expected to have higher values compared to normal subjects.

Features belong to vector $\vec{PV}$ assumed low values in depressed voices compared to normal, the opposite of that, features from $\vec{AP}$, low values means less pauses time which are expected for normal and higher values in depression condition.

More interesting features belong to the last two vectors $\vec{CV}$ and $\vec{CP}$. Features belong to $\vec{CV}$ measure all voice segments time that are above the mean time of all pauses, for example, "T_{time}_CV" compute the total time for all voice segments generated by subject and were greater than his average pause time. The values from these features expected to be larger for normal subjects compared to patients suffer from depression. The reverse of that is $\vec{CP}$ vector features, which tell if someone is speaking with more pauses than his average of utterances.

Based on these features, extra statistical and ratio functions can be derived, which may also contribute in discriminating depressed voices, for example:

$$R_{no}\_AV = \frac{No\_AV}{No\_V} \tag{7.25}$$

$$R_{Time}\_AV = \frac{No\_AV}{T_{time}\_V} \tag{7.26}$$

$$R_{mean}\_AV = \frac{Mean\_V}{Mean\_AV} \tag{7.27}$$

$$R_{var}\_AV = \frac{VAR\_V}{VAR\_AV} \tag{7.28}$$

$$Mean_{dif}\_AV = \mid Mean\_AV - Mean\_V \mid \tag{7.29}$$

$$STD_{dif}\_AV = \mid STD\_AV - STD\_V \mid \tag{7.30}$$

$$VAR_{dif}\_AV = \mid VAR\_AV - VAR\_V \mid \tag{7.31}$$

$$R_{(mean,no)}\_AV = \frac{R_{mean}\_AV}{R_{no}\_V} \tag{7.32}$$

$$R_{no}\_PP = \frac{No\_PP}{No\_P} \tag{7.33}$$

$$R_{mean}\_PP = \frac{Mean\_P}{Mean\_PP} \tag{7.34}$$

$$R_{var}\_PP = \frac{VAR\_P}{VAR\_PP} \tag{7.35}$$

$$Mean_{dif}\_PP = \mid Mean\_PP - Mean\_P \mid \tag{7.36}$$

$$STD_{dif}\_PP = \mid STD\_PP - STD\_P \mid \tag{7.37}$$

$$VAR_{dif}\_PP = \mid VAR\_PP - VAR\_P \mid \tag{7.38}$$

$$R_{(mean,no)}\_PP = \frac{R_{mean}\_PP}{R_{no}\_PP} \tag{7.39}$$

$$R_{mean}\_(AV, PP) = \frac{Mean\_AV}{Mean\_PP} \tag{7.40}$$

$$Mean_{dif}\_(AV, PP) = \mid Mean\_AV - Mean\_PP \mid \tag{7.41}$$

$$Tmean_{time}\_(AV, PP) = Mean\_AV + Mean\_PP \tag{7.42}$$

$$R_{mean(AV,PP)}\_(VP) = \frac{Tmean_{time}\_(AV, PP)}{T_{time}\_VP} \tag{7.43}$$

$$T_{No}\_(AV, PP) = No\_AV + No\_PP \tag{7.44}$$

$$No_{dif}\_(AV, PP) = \mid No\_AV - No\_PP \mid \tag{7.45}$$

$$R_{No}\_(AV, PP) = \frac{No\_AV}{No\_PP} \tag{7.46}$$

$$T_{time}\_(AV, PP) = T_{time}\_AV + T_{time}\_PP \tag{7.47}$$

$$T_{dif}\_(AV, PP) = \mid T_{time}\_AV - T_{time}\_PP \mid \tag{7.48}$$

$$R_{time}\_(AV, PP) = \frac{T_{time}\_AV}{T_{time}\_PP} \tag{7.49}$$

$$R_{T(AV,PP)}\_(VP) = \frac{T_{time}\_(AV, PP)}{T_{time}\_VP} \tag{7.50}$$

$$R_{(time,No)}\_(AV, PP) = \frac{T_{time}\_(AV, PP)}{T_{No}\_(AV, PP)} \tag{7.51}$$

$$R_{(T\_mean,No\_dif)}\_(AV, PP) = \frac{Tmean_{time}\_(AV, PP)}{No_{dif}\_(AV, PP)} \tag{7.52}$$

$$R_{(T\_mean,R\_No)}\_(AV,PP) = \frac{Tmean_{time}\_(AV,PP)}{R_{No}\_(AV,PP)} \tag{7.53}$$

$$R_{(T\_dif,No\_dif)}\_(AV,PP) = \frac{T_{dif}\_(AV,PP)}{No_{dif}\_(AV,PP)} \tag{7.54}$$

$$R_{(T\_dif,R\_No)}\_(AV,PP) = \frac{T_{dif}\_(AV,PP)}{R_{No}\_(AV,PP)} \tag{7.55}$$

$$R_{(R\_time,No\_dif)}\_(AV,PP) = \frac{R_{time}\_(AV,PP)}{No_{dif}\_(AV,PP)} \tag{7.56}$$

$$R_{(R\_time,R\_No)}\_(AV,PP) = \frac{R_{time}\_(AV,PP)}{R_{No}\_(AV,PP)} \tag{7.57}$$

$$R_{no}\_CV = \frac{No\_CV}{No\_V} \tag{7.58}$$

$$R_{mean}\_CV = \frac{Mean\_V}{Mean\_CV} \tag{7.59}$$

$$R_{var}\_CV = \frac{VAR\_V}{VAR\_CV} \tag{7.60}$$

$$Mean_{dif}\_CV = \mid Mean\_CV - Mean\_V \mid \tag{7.61}$$

$$STD_{dif}\_CV = \mid STD\_CV - STD\_V \mid \tag{7.62}$$

$$VAR_{dif}\_CV = \mid VAR\_CV - VAR\_V \mid \tag{7.63}$$

$$R_{(mean,no)}\_CV = \frac{R_{mean}\_CV}{R_{no}\_V} \tag{7.64}$$

$$R_{(mean,no)}\_CV = \frac{R_{mean}\_CV}{R_{no}\_CV} \tag{7.65}$$

$$R_{no}\_CP = \frac{No\_CP}{No\_P} \tag{7.66}$$

$$R_{mean}\_CP = \frac{Mean\_P}{Mean\_CP} \tag{7.67}$$

$$R_{var}\_CP = \frac{VAR\_P}{VAR\_CP} \tag{7.68}$$

$$Mean_{dif}\_CP = \mid Mean\_CP - Mean\_P \mid \tag{7.69}$$

$$STD_{dif}\_CP = \mid STD\_CP - STD\_P \mid \tag{7.70}$$

$$VAR_{dif}\_CP = \mid VAR\_CP - VAR\_P \mid \tag{7.71}$$

$$R_{(mean,no)}\_CP = \frac{R_{mean}\_CP}{R_{no}\_PP} \tag{7.72}$$

$$R_{mean}\_(CV, CP) = \frac{Mean\_CV}{Mean\_CP} \tag{7.73}$$

$$Mean_{dif}\_(CV, CP) = \mid Mean\_CV - Mean\_CP \mid \tag{7.74}$$

$$Tmean_{time}\_(CV, CP) = Mean\_CV + Mean\_CP \tag{7.75}$$

$$R_{mean(CV,CP)}\_(VP) = \frac{Tmean_{time}\_(CV, CP)}{T_{time}\_VP} \tag{7.76}$$

$$T_{No}\_(CV, CP) = No\_CV + No\_CP \tag{7.77}$$

$$No_{dif}\_(CV, CP) = \mid No\_CV - No\_CP \mid \tag{7.78}$$

$$R_{No}\_(CV, CP) = \frac{No\_CV}{No\_CP} \tag{7.79}$$

$$T_{time}\_(CV, CP) = T_{time}\_CV + T_{time}\_CP \tag{7.80}$$

$$T_{dif}\_(CV, CP) = \mid T_{time}\_CV - T_{time}\_CP \mid \tag{7.81}$$

$$R_{time}\_(CV, CP) = \frac{T_{time}\_CV}{T_{time}\_CP} \tag{7.82}$$

$$R_{T(CV,CP)}\_(VP) = \frac{T_{time}\_(CV, CP)}{T_{time}\_VP} \tag{7.83}$$

$$R_{(time,No)}\_(CV, CP) = \frac{T_{time}\_(CV, CP)}{T_{No}\_(CV, CP)} \tag{7.84}$$

$$R_{(T\_mean,No\_dif)}\_(CV, CP) = \frac{Tmean_{time}\_(CV, PP)}{No_{dif}\_(CV, CP)} \tag{7.85}$$

$$R_{(T\_mean,R\_No)}\_(CV, CP) = \frac{Tmean_{time}\_(CV, CP)}{R_{No}\_(CV, CP)} \tag{7.86}$$

$$R_{(T\_dif,No\_dif)}\_(CV, CP) = \frac{T_{dif}\_(CV, CP)}{No_{dif}\_(CV, CP)} \tag{7.87}$$

$$R_{(T\_dif,R\_No)}\_(CV, CP) = \frac{T_{dif}\_(CV, CP)}{R_{No}\_(CV, CP)} \tag{7.88}$$

$$R_{(R\_time,No\_dif)}\_(CV,CP) = \frac{R_{time}\_(CV,PP)}{No_{dif}\_(CV,CP)} \tag{7.89}$$

$$R_{(R\_time,R\_No)}\_(CV,CP) = \frac{R_{time}\_(CV,CP)}{R_{No}\_(CV,CP)} \tag{7.90}$$

$$R_{No(AP,AV)}\_(VP) = \frac{No\_AP + No\_AV}{T_{No}\_VP} \tag{7.91}$$

$$R_{T(AP,AV)}\_(VP) = \frac{T_{time}\_AP + T_{time}\_AV}{Mean\_VP} \tag{7.92}$$

$$R_{Mean(AP,AV)}\_(VP) = \frac{Mean\_AP + Mean\_AV}{T_{time}\_VP} \tag{7.93}$$

$$R_{No(PV,PP)}\_(VP) = \frac{No\_PV + No\_PP}{T_{No}\_VP} \tag{7.94}$$

$$R_{T(PV,PP)}\_(VP) = \frac{T_{time}\_PV + T_{time}\_PP}{Mean\_VP} \tag{7.95}$$

$$R_{Mean(PV,PP)}\_(VP) = \frac{Mean\_PV + Mean\_PP}{T_{time}\_VP} \tag{7.96}$$

$$R_{no}\_(AP,PP) = \frac{No\_AP}{No\_PP} \tag{7.97}$$

$$R_{mean}\_(AP,PP) = \frac{Mean\_AP}{Mean\_PP} \tag{7.98}$$

$$R_{time}\_(AP,PP) = \frac{T_{time}\_AP}{T_{time}\_PP} \tag{7.99}$$

$$R_{no}\_(PV,AV) = \frac{No\_PV}{No\_AV} \tag{7.100}$$

$$R_{mean}\_(PV,AV) = \frac{Mean\_PV}{Mean\_AV} \tag{7.101}$$

$$R_{time}\_(PV,AV) = \frac{T_{time}\_PV}{T_{time}\_AV} \tag{7.102}$$

$$R_{time}\_(AV,AP) = \frac{T_{time}\_AV}{T_{time}\_AP} \tag{7.103}$$

Table 7.2 Final set of acoustic features.

Features	Functions	No. of features
Pitch	Mean,var,std	3
Harmonic-to-noise ratio	Mean,var,std	3
Noise-to-harmonic ratio	Mean,var,std	3
Shimmer (six scales)	Mean,var,std	18
Jitter (five scales)	Mean,var,std	15
Number of voice breaks	Mean,var,std	3
Fractions of locally unvoiced frames	Mean,var,std	3
Degree of voice breaks	Mean,var,std	3
Number of voice breaks	Mean,var,std	3
Number of periods	Mean,var,std	3
Auto_correlation	Mean,var,std	3
Number of pulses	Mean,var,std	3
MFCC	Min,max,skewness,kurtosis,mean,var,std	7
Fbank)	Min,max,skewness,kurtosis,mean,var,std	7
SSC	Min,max,skewness,kurtosis,mean,var,std	7
Intensity	Mean,var,std	3
F1(Hz)	Mean,var,std	3
Intensity at F1	Mean,var,std	3
B1(Hz) bandwidth_F1	Mean,var,std	3
F2(Hz)	Mean,var,std	3
Intensity at F2	Mean,var,std	3
B2(Hz) bandwidth_F2	Mean,var,std	3
ZCR	Mean,var,std	3
Energy	Mean,var,std	3
Energy_entropy	Mean,var,std	3
Spectral_centroid	Mean,var,std	3
Spectral_spread	Mean,var,std	3
Spectral_flux	Mean,var,std	3
Spectral_rolloff	Mean,var,std	3
Spectral_entropy	Mean,var,std	3
New developed features	Mean,var,std, ratio, difference	99
Total		228

$$R_{time\_}(PP,PV) = \frac{T_{time\_}PP}{T_{time\_}PV} \tag{7.104}$$

These features are an additional measurement of speaking behaviour tend to capture depression effect on speech and they will be investigated in the following sections for their usefulness in detecting depression.

In addition to the above mentioned features, also features from the previous chapters 5 and 6 will be included. Table 7.2 list all features used in our proposed depression evaluation system.

Table 7.3 AVEC-2013 t-test statistical significance. STD: Standard deviation, VAR: Variance, DF: Degree of freedom,class(0) n=26, class(1) n =24.

Rank	Features	t-test	DF	P-Value
1	Mean-mean-harmonics-to-noise-ratio	3.663	48	0.001
2	Mean-minimum-pitch	3.492	48	0.001
3	Mean-mean-pitch	3.249	48	0.002
4	Mean-median-pitch	3.244	48	0.002
5	Mean-mean-autocorrelation	3.045	48	0.004
6	VAR_spectral_spread	2.82	48	0.007
7	STD_spectral_spread	2.808	48	0.007
8	Mean_Pitch	2.755	48	0.008
9	Mean-turn-length	2.675	48	0.01
10	Mean_B1Hz	-2.649	48	0.011
11	Mean-maximum-pitch	2.622	48	0.012
12	$\mathbf{R}_{Mean(AP,AV)\_}(VP)$	-2.608	48	0.012
13	VAR-mean-period	-2.443	48	0.018
14	$\mathbf{R}_{mean(CV,CP)\_}(VP)$	-2.417	48	0.02
15	STD-mean-period	-2.3	48	0.026
16	STD_FBANK	-2.271	48	0.028
17	STD-turn-length	2.263	48	0.028
18	VAR_FBANK	-2.255	48	0.029
19	$\mathbf{R}_{T(AP,AV)\_}(VP)$	2.216	48	0.031
20	$\mathbf{T}_{No\_}(CV,CP)$	2.13	48	0.038
21	$\mathbf{R}_{(time,No)\_}(CV,CP)$	-2.055	48	0.045
22	STD_B1Hz	-2.05	48	0.046
23	$\mathbf{R}_{(mean,no)\_}PP$	2.027	48	0.048
24	$\mathbf{R}_{time\_}(AP,PP)$	2.027	48	0.048

7.2.2 Statistical analysis

7.2.2.1 AVEC2013

Table 7.3 list the results of the independent sample t-test, and only for features that have a 95% significance level in mean differences between healthy and depressed participants. The comparison shows that seven variables from the newly derived set have significant mean difference and Fig 7.3 shows the distribution of these features cross the two compared groups. The feature "$\mathrm{R}_{Mean(AP,AV)\_}(VP)$" which is the ratio of sum of the mean time of both active voice and active pause segments to the total recording time.

This indicate if some one speaking more than his average and he/she produce pauses less than his/her average, then he/she will likely to be normal and not depressed. According to this

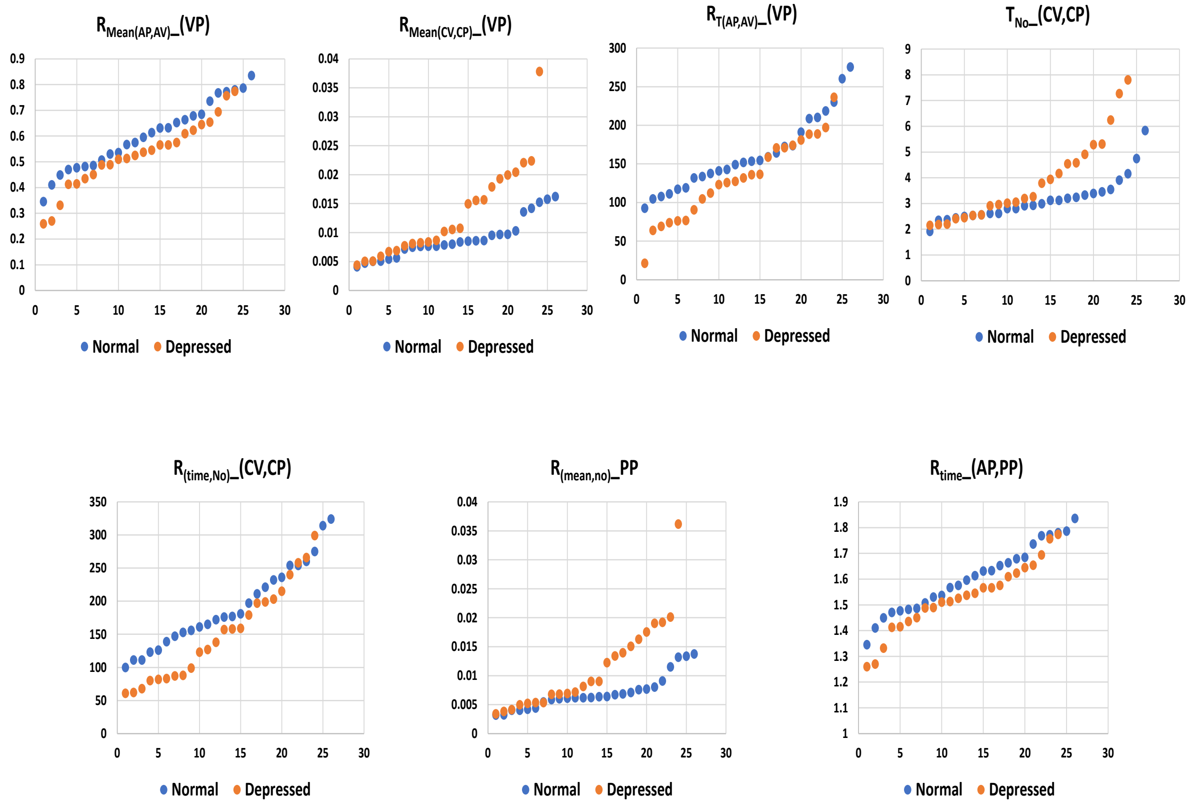


Figure 7.3 AVEC-2013 the distributions of the new most significance features based on the t-test.

variable, the average value for the normal group was 0.6 compared to 0.5 for depressed group.

The "R$_{mean(CV,CP)\_}(VP)$" measures the ratio of the sum for the mean for both cross voice and cross pause segments to the total time. This means if someone produce more pauses than his average utterances and speaks more than his average pauses is likely to be depressed. The mean value for the depressed class was 0.03 while the mean for normal subjects was 0.008.

The third significant feature is "R$_{T(AP,AV)\_}(VP)$", which estimate the ratio of the sum for both total times for active voice and active pause segments time to the mean of both voices and pauses time. The high value for this feature indicate normal state, it describes that the individual utterances above his mean speech combined with his pauses time below his average pauses time. The normal group ratio was 161.9 compared to 130.6 for the patients group.

The fourth important feature was "T$_{No\_}(CV,CP)$" this is the total number of all cross pause and cross voice segments. The higher value in this dataset indicate depression state. For this feature, the average number of segments for the depressed group was 3.86 compared to 3.13 for normal subjects. However, the R$_{(time,No)}\_(CV,CP)$ feature, which is dividing the total time for cross voices and cross pauses segments time by "T$_{No}\_(CV,CP)$" will indicate the opposite, larger number means no depression. The healthy subjects average value was 191.3 compared to 151.1 for patients class.

The "R$_{(mean,no)}\_PP$" feature compute the ratio of average passive pauses time to the number of passive pauses segments. Larger value may refer as depression condition. The average value for this feature was 0.007 for the normal class, and 0.011 for depressed patients.

The last significant feature is "R$_{time}\_(AP,PP)$" compute the ratio of the total time for active pause segments to the total time for passive voice segments. The high value was for the healthy subjects at 1.6 compared to 1.5 for the patients group, this might reflect shorter pauses even though voice segments may no be the dominant segments.

This statistical results indicate that these features have useful information and could be used in characterising behaviour in speech associated with depression. However, this needs further investigation when performing the feature selection step in the model evaluation, as this step will select features based on their contribution to the classification accuracy, thus other features maybe selected and/or overlapped compared to the two approaches.

7.2.2.2 AVEC2014

The t-test results for the AVEC-2014 were listed in Table 7.4. Although more samples are available in AVEC-2014 compare to AVEC-2013, only two variables from the newly developed features had a significance mean difference, whereas in AVEC-2013 seven from the new feature set had statistical mean significance. Fig 7.4 shows the distribution of these new features with significance value. This may due to the fact that the AVEC-2014 verbal tasks were shorter in times (mean=52.6 sec) compared to AVEC-2013 (mean= 845.6), thus shorter speech may not provide enough time for the new features to capture a behaviour. The "Mean$_{dif}\_AV$" feature measure the absolute difference between the mean of active voice segments time and the average voice segments time. Larger value indicate the subject is an active speaker and produce much larger utterances compare to his average voice segments. The normal group have average value of 0.49 compared to 0.34 for depression patients. The last feature is "R$_{mean}\_CP$" which compute the ratio of mean pause time to the average cross pauses time. This variable estimate if

Table 7.4 AVEC-2014 t-test statistical significance. Class(0) n=52, class(1) n =48

Rank	**Features**	**t-test**	**DF**	**P-Value**
1	VAR_spectral_spread	-3.26	98	0.002
2	STD_spectral_spread	-3.12	95.8	0.002
3	STD_FBANK	2.53	75.7	0.013
4	Mean_F2Hz	2.47	98	0.015
5	Mean_number_of_pulses	-2.29	98	0.024
6	Mean_number_of_periods	-2.28	98	0.024
7	VAR_BANK	2.56	98	0.026
8	std_F1Hz	2.2	97.7	0.03
9	Mean_B1Hz	2.05	97.9	0.042
10	**Mean**$_{dif}\_AV$	-2.01	84.9	0.047
11	**R**$_{mean}\_CP$	-2.02	61.16	0.047
12	Mean_maximum_pitch	-1.99	98	0.049

someone had mean pause time higher than value. This value represents the mean of only pauses that are above his average voice segments, this person probably have no depression. The mean for normal group was 0.37 and 0.16 for patients group.

7.2.2.3 DAIC-WOZ

The t-test features's significance for the DAIC-WOZ dataset are listed in Table 7.5, and only twelves features that have a significance mean difference between normal/depressed subjects. Similar to AVEC-2013, seven features (58% from the whole list) were from the new derived feature set. Fig 7.5 shows the distribution of the new seven features cross the normal and depressed classes.

The first significant feature is R$_{var}\_PP$ which is estimate the ratio of variance time for mean pause segments to the variance of passive pause segment times. Class 0 have smaller mean of 0.77 compare to patients group mean ratio of 0.99. The second "VAR$_{dif}\_PP$" and the third feature "STD$_{dif}\_PP$" measuring the absolute difference value between the variance and the standard deviation respectively for pauses and passive pause segments time. In contrast to the first feature, higher values were for the normal group at 0.46, 0.18 respectively for the second and third features compared to 0.12,0.09 for the patients. The forth and the fifth significant features are "R$_{(mean,no)}\_PP$" and "R$_{time}\_(AP,PP)$" which both were also significant in AVEC-2013 datsets, the normal group have mean value of 1.89,0.89 for the "R$_{(mean,no)}\_PP$" and "R$_{time}\_(AP,PP)$" respectively, whereas patients group have mean values of 1.78 fourth

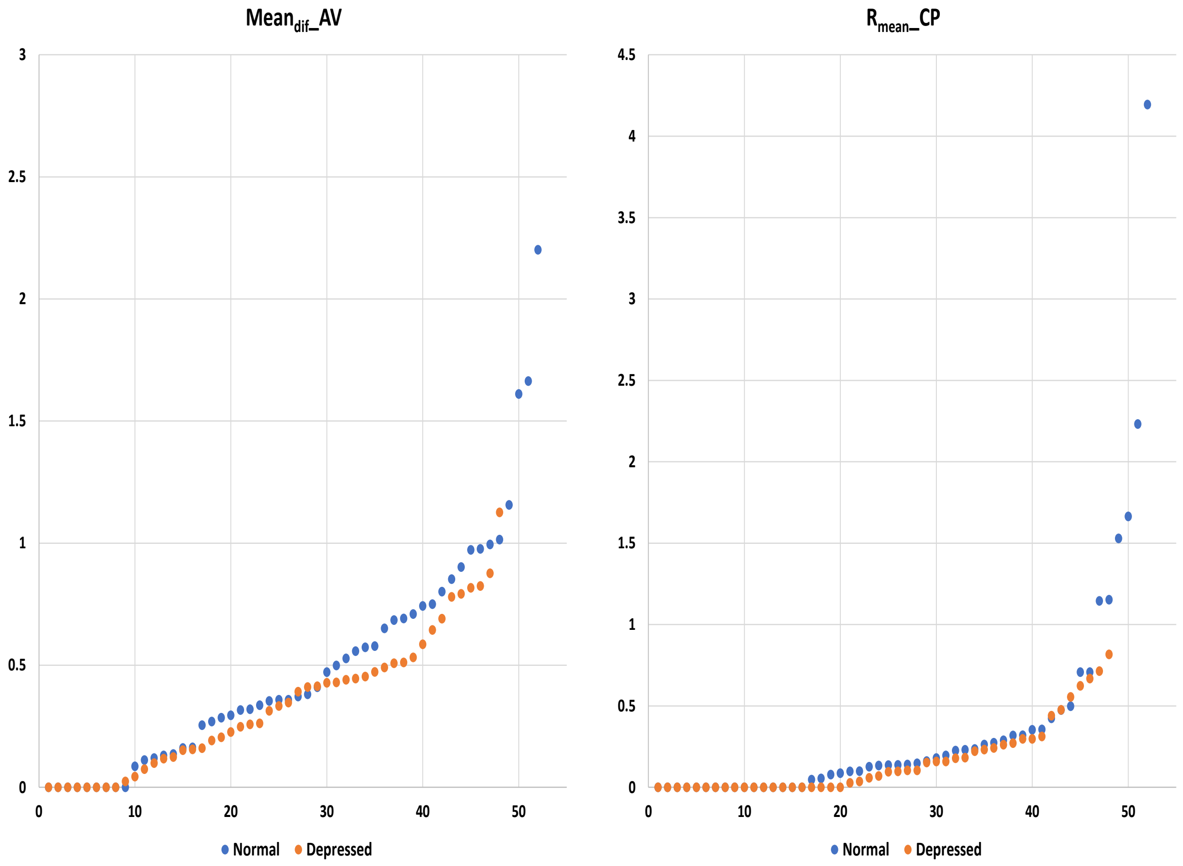


Figure 7.4 AVEC-2014 the distributions of the new most significance features based on the t-test.

Table 7.5 DAIC-WOZ t-test significance features. Class(0) n=77, class(1) n =30.

Rank	Features	t-test	DF	P-Value
1	Mean-Number-of-voice-breaks	3.061	102.105	0.003
2	$\mathbf{R}_{var}\_PP$	-2.668	105	0.009
3	$\mathbf{VAR}_{dif}\_PP$	2.398	89.359	0.019
4	Mean-Number-of-pulses	2.353	96.361	0.021
5	Mean-Number-of-periods	2.324	95.945	0.022
6	$\mathbf{STD}_{dif}\_PP$	2.315	104.938	0.023
7	Kort_FBANK	-2.331	37.410	0.025
8	$\mathbf{R}_{(mean,no)}\_PP$	2.207	105	0.029
9	$\mathbf{R}_{time}\_(AP,PP)$	2.207	105	0.029
10	$\mathbf{R}_{var}\_CV$	-2.098	105	0.038
11	$\mathbf{R}_{no}\_(AP,PP)$	2.084	98.104	0.040
12	STD- Number- of- voice-breaks	2.030	73.969	0.046

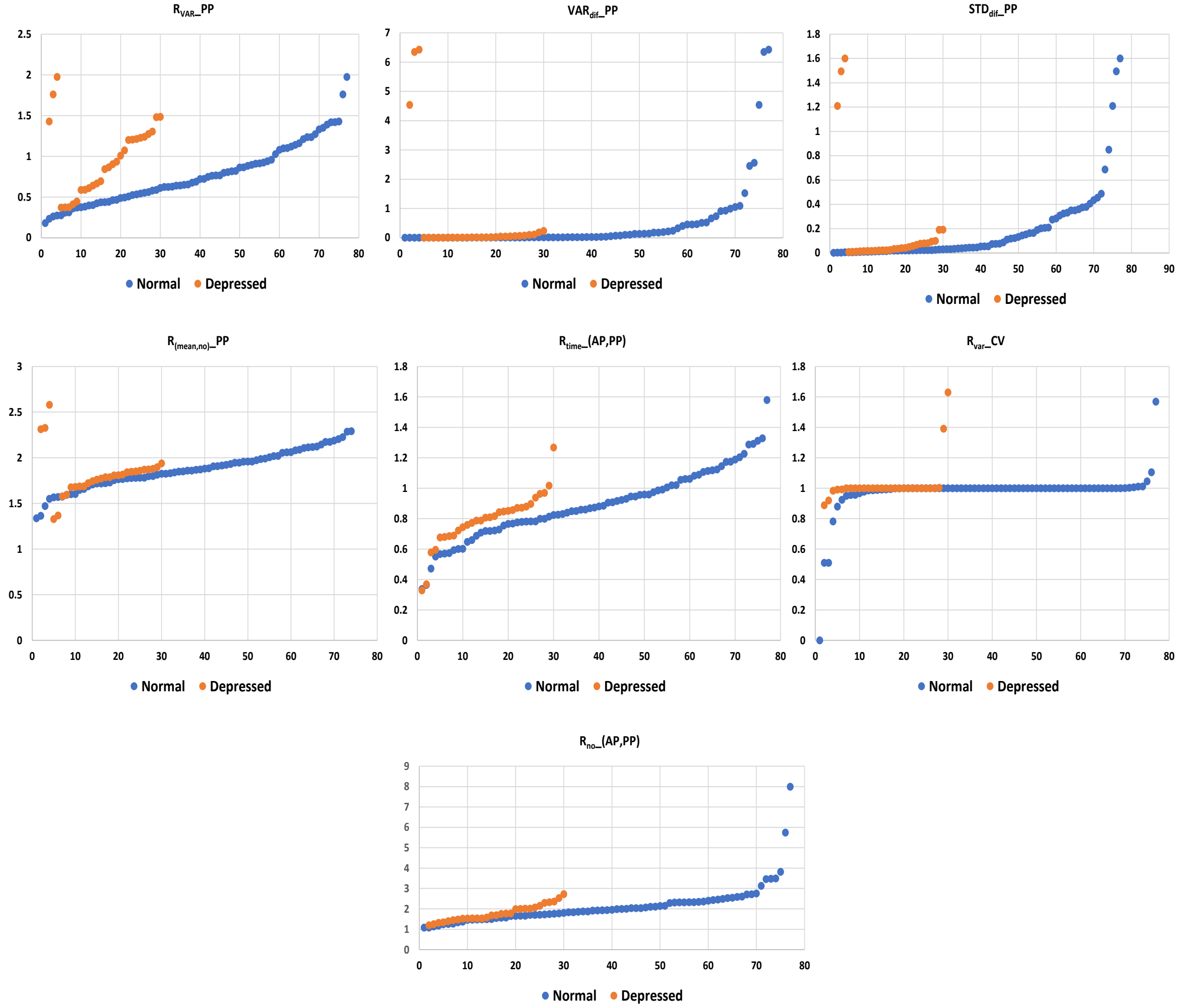


Figure 7.5 AVEC-2016 the distributions of the new most significance features based on the t-test.

and 0.78 fifth features.

The "R_{var}_CV" is the sixth important features. It is the ratio of time variance for voice segments to variance of cross voice segments time. Normal subjects mean value was 0.9 compared to 1.15 for depression patients. The last significant feature from the new developed variable is "$\mathrm{R}_{no}\_(AP, PP)$". This feature estimate the ratio of the total number of active pause segments to the total number of passive pause segments. Higher value found in healthy group with mean of 2.1 compared to 1.8 for depression patients.

7.2.3 Results

7.2.3.1 AVEC-2013

Two approaches were adopted to perform the depression evaluation task. The first approach is an audio-based regression model aims to predict the clinical test scores associated with each speaker's recording, and the second model is an audio-based classifier that identifies those speakers that suffer from depression from others with minimal or no depression. The three data partitions from the AVEC-2013 were used to build the two models. The model is fitted with samples from the training set and evaluated two times. Firstly, against samples from the development set and secondly, with recordings from the test set and this scenario was iterated several times. Each time model is optimised with a set of parameters known as the tuning parameters. Until exhausting all those variables, the best configuration of tuning variables which provide the maximum performance is selected. The GBC and GBR algorithms [256] were used to build the two approaches, and the built-in feature importance was utilised to help select the most informative features. Table 7.6 lists the most informative features.

The performances of the classification and regression models listed in Table 7.7. The proposed model outperformed both modalities of the baseline performances in predicting BDI-scores. The baseline audio model have (MAE, RMSE) of (8.6, 10.7), and (10.3, 14.1) estimated for the development and test sets respectively, while the proposed model predicting (MAE, RMSE) errors at (6.6, 8.9) Fig and (6.8, 8.7) for the development and test sets respectively. The proposed model performance was also better than the baseline video model which have (MAE, RMSE) of (8.7,10.7), and (10.8,13.1) for development and test sets respectively. Another design also tested, in which the the samples for both training and development sets were concatenated

Table 7.6 AVEC-2013 ranked features

Rank	Significant features	Weight	Rank	Significant features	Weight
1	Mean -Mean_HNR	0.024368	**29**	RT(PV,PP)_(V P)	0.005889
2	Mean -Median pitch	0.017486	30	STD -Jitter (local_absolute)	0.005824
3	Mean -Mean pitch	0.015462	31	Mean -Jitter (ppq5)	0.005682
4	Mean_Pitch	0.014202	**32**	No_P	0.005358
5	Mean -Minimum pitch	0.013474	33	var_spectral_spread	0.00534
6	Mean -Mean autocorrelation	0.012871	**34**	T_(No)_(AV,PP)	0.005325
7	T_No_(CV,CP)	0.010908	35	std_spectral_rolloff	0.005269
8	Mean turn length	0.00988	**36**	VAR_V	0.004964
9	R_(Mean(AP,AV)_(VP)	0.009055	37	Var -Mean_ HNR	0.004919
10	STD -Minimum pitch	0.008613	**38**	T_(time}_(CV,CP)	0.004911
11	STD -Median pitch	0.008543	**39**	No_V	0.00491
12	STD turn length	0.008277	**40**	R_(R_time,R_No)_(CV,CP)	0.004907
13	Var -Mean period	0.008213	41	var_spectral_entropy	0.004906
14	std_FBANK	0.007783	**42**	VAR_(dif)_PP	0.004835
15	Var -Minimum pitch	0.007493	**43**	No_AP	0.004734
16	Var turn length	0.007415	**44**	Mean_(dif)_PP	0.004634
17	std_spectral_entropy	0.007176	**45**	R_(time,No)_(CV,CP)	0.004626
18	std_spectral_spread	0.007105	46	Mean_B1(Hz)	0.004618
19	Mean -Jitter (local)	0.007049	47	STD -Mean autocorrelation	0.004507
20	R_mean(AV,PP)_(V P)	0.007008	48	Mean -Mean _NHR	0.004499
21	Var -Median pitch	0.007002	49	Var -Mean autocorrelation	0.004423
22	STD -Mean period	0.006922	50	Mean_energy_entropy	0.004331
23	VAR -Mean NHR	0.006835	**51**	Mean_P	0.004323
24	P_(max)R_T(VP)	0.006618	52	Var -Number of pulses	0.004285
25	No_PP	0.006398	**53**	R_(time)_(AP,PP)	0.004259
26	T_(time)_AP	0.006397	**54**	VAR_PP	0.004246
27	STD -Shimmer (apq5)	0.006186	55	Min_SSC	0.004184
28	Var_FBANK	0.006183	56	Mean -Fraction of locally unvoiced frames	0.004159

and used to fit the model. Likewise, the model optimised and evaluated as explained above, a better results were achieved, as the model (MAE, RMSE) prediction errors dropped to 6.2 and 8.4, respectively. This improvement was expected, due to the fact that model trained with larger samples 100 vs 50 compared to the previous configuration, so the model learned more from the extra number of samples and therefore performed better.

Using only the new developed features was also investigate, and the results of the proposed model are shown in Fig 7.6. The results were also better than both of the baseline models, the MAE= 7.4, RMSE=10.1 for the development, and MAE=7.8 and RMSE= 10.06 for the test set, this indicate that these features have considerable impact on the performance of the model.

Table 7.7 AVEC-2013 full results comparison with baseline.

Method	Modality	Train test	Test set	MAE	RMSE	ACC.%	F1_Score	Precision	Recall	Confusion Matrix	
Baseline [245]	**Audio**	Train	Dev	**8.66**	10.75	N/A	N/A	N/A	N/A	N/A	
	Video	Train	Dev	8.74	10.72	N/A	N/A	N/A	N/A	N/A	
	Audio	Train	Test	**10.35**	14.12	N/A	N/A	N/A	N/A	N/A	
	Video	Train	Test	10.88	13.16	N/A	N/A	N/A	N/A	N/A	
Williamson [179]	Audio	Train	Test	5.75	7.4	N/A	N/A	N/A	N/A	N/A	
Meng [180]	Audio	Train	Dev	9.7	11.5	N/A	N/A	N/A	N/A	N/A	
	Audio-Video	Train	Dev	6.94	8.54	N/A	N/A	N/A	N/A	N/A	
	Audio-Video	Train	Test	8.72	10.96	N/A	N/A	N/A	N/A	N/A	
Cummins [182]	Audio	Train	Dev	N/A	10.44	N/A	N/A	N/A	N/A	N/A	
	Audio	Train	Test	N/A	10.17	N/A	N/A	N/A	N/A	N/A	
Cummins [183]	Audio	Train	Dev	N/A	7.4	N/A	N/A	N/A	N/A	N/A	
	Audio	Train	Test	N/A	9.49	N/A	N/A	N/A	N/A	N/A	
Zhu [197]	Video	Train	Test	7.58	9.82	N/A	N/A	N/A	N/A	N/A	
Kaya [184]	Audio-Video	Tain	Test	7.84	10.22	N/A	N/A	N/A	N/A	N/A	
Kachele [186]	Audio	Train	Dev	9.35	11.4	N/A	N/A	N/A	N/A	N/A	
	Video	Train	Dev	7.03	8.82	N/A	N/A	N/A	N/A	N/A	
	Audio-Video	Train	Dev	8.3	9.94	N/A	N/A	N/A	N/A	N/A	
	Audio	Train	Test	10.35	14.12	N/A	N/A	N/A	N/A	N/A	
	Video	Train	Test	8.97	10.82	N/A	N/A	N/A	N/A	N/A	
	Audio-Video	Train	Test	8.72	10.96	N/A	N/A	N/A	N/A	N/A	
Proposed	**Audio**	Train	Dev	**6.6**	**8.9**	84	0.84 (0.84/0.84)	0.84 (0.88/0.81)	0.84 (0.81/0.88)	**21**	5
										3	**21**
	Audio	Train	Test	**6.8**	**8.7**	82	0.82 (0.82/0.82)	0.82 (0.83/0.81)	0.82 (0.84/0.80)	**20**	5
										4	**21**
	Audio	Train+Dev.	Test	**6.2**	**8.4**	86	0.86 (0.87/0.84)	0.88 (0.80/0.95)	0.86 (0.96/0.76)	**24**	1
										6	**19**

The second model that was proposed is a binary classification procedure, in which the aim was to discriminate participant's class, either class 0 (i.e. no/or minimal depression) or 1 depressed condition. Unfortunately, there were no baseline performances for this task as the aim of the challenge was BDI scores prediction. The proposed classification model achieved accuracy of 84% and 82% computed for the development and test sets respectively; however, this accuracy increased to 86% when fitting the model with a combination of training and development data. Further, other evaluation metrics scores were estimated, including the F1_scores, precision and recall per each class and as well as weighted average score, and listed in Table 7.7. The sensitivity and specificity also computed for the three configurations (test vs development, test and both combined) and obtained scores of (84%,87%), (80%, 84%) and (96%, 76%) respectively. Finally the confusion matrix was estimated (the values in blue refer to the correct normal class prediction, the values in green are the correct depression class prediction, while the red values represents all missed classification)

7.2.3.2 AVEC2014

Similar to AVEC-2013, the the same evaluation approach was used to compute the performance of the two models classification and regression and Table 7.8 shows the most significant fea-

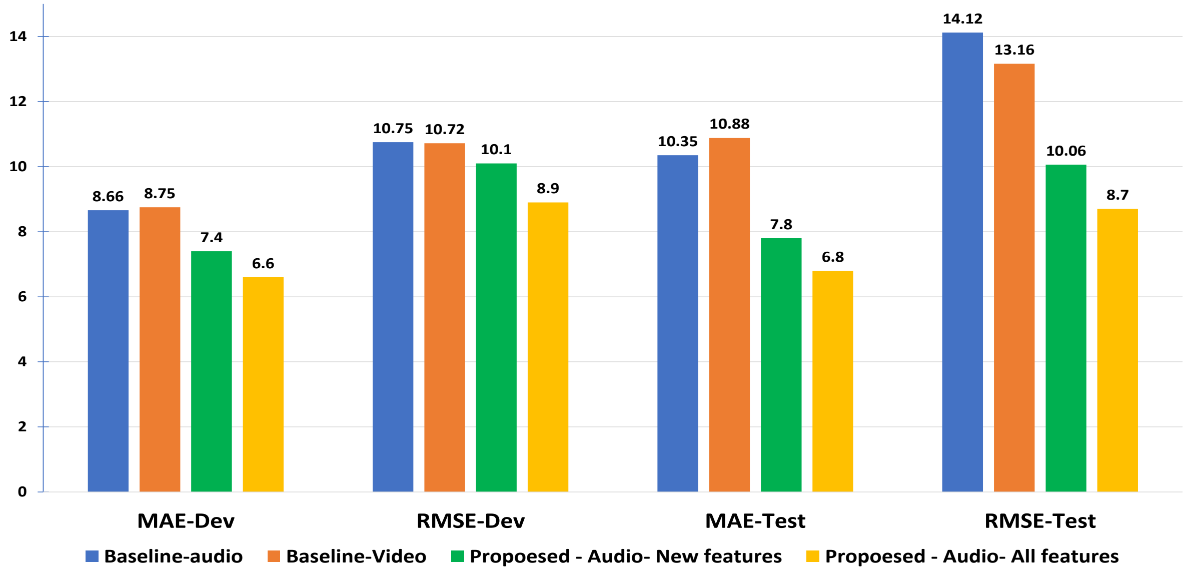


Figure 7.6 AVEC-2013 results comparison with model built using only the new features

tures. However, with this dataset there are two short-verbal tasks the NW and the FF, and base on that, several approaches were investigated includes using them as separate or combined to build and evaluated the proposed models. This approach will also shows which verbal task is more informative in detection and estimating depression severity. The results listed in Table 7.9.

The baseline results for the audio model were MAE=8.9, RMSE=11.52 for the development set and MAE=10.03, RMSE=12.56 for the test set. The proposed model achieved better results, the MAE, RMSE were 7.1,9.4 for the development set and 7.1,9.4 for the test set. In fact the developed model even achieved better results than the baseline video and combined audio-video modalities except for the combined development model which were better than the proposed model with MAE, RMSE scores at 6.68,8.34 respectively. However, training and testing using the NW and FF tasks separately, the developed model outperformed all baseline modalities including the combined audio-video, the NW task model achieved better scores at 6.5,8.7 (MAE,RMSE) for the development and 5.8,8.2 for the test, while FF task-base model scores were 6.8,97 and 6.5,9.8 for the development and test sets respectively. A slightly better results were produced when combined both the training and development samples and evaluate the model performance, in this case the model achieved MAE = 7.26 and RMSE = 9.85, this improvement because the model trained with 200 samples instead of 100 samples from just the

Table 7.8 AVEC-2014 ranked features

Rank	Significant features	Weight	Rank	Significant features	Weight
1	Mean_B1(Hz)	0.000913	34	Skew_MFCC	0.000239
2	Mean_energy_entropy	0.000804	35	Mean_spectral_spread	0.000239
3	Mean_T_intensity	0.000788	36	Mean_F1(Hz)	0.000238
4	Mean -Mean period	0.000787	37	STD -Shimmer (local_ dB)	0.000234
5	Mean_MFCC	0.000774	38	STD_SSC	0.000229
6	Mean_SSC	0.000752	**39**	No_(dif)_(AV,PP)	0.000215
7	R(mean,no)_CV	0.000618	40	STD -Standard deviation	0.000205
8	STD_I2	0.0006	41	Mean_energy	0.000199
9	R_(R_time,No_dif)_(CV,CP)	0.000596	42	Kort_MFCC	0.000198
10	Mean_I1	0.000583	43	VAR -Mean NHR	0.000197
11	R_(var)_AV	0.000573	44	VAR_MFCC	0.000195
12	VAR -Standard deviation of period	0.000562	**45**	Mean_P	0.000192
13	VAR_FBANK	0.000488	46	VAR -Standard deviation	0.000186
14	STD_T_intensity	0.000472	47	STD_spectral_flux	0.000179
15	Mean -Shimmer (apq5)	0.000449	48	Mean -Degree of voice breaks	0.000178
16	STD_FBANK	0.000449	49	VAR_spectral_entropy	0.000166
17	Mean_Pitch	0.000426	50	VAR_spectral_flux	0.000165
18	Mean -Mean autocorrelation	0.000409	**51**	STD_(dif)_PP	0.000161
19	STD -Shimmer (apq5)	0.000402	52	Skew_FBANK	0.00016
20	VAR_B2(Hz)	0.000399	53	Kort_FBANK	0.000159
21	Min_SSC	0.000396	54	Mean_AP	0.000159
22	VAR_T_intensity	0.000391	55	VAR -Mean period	0.000158
23	Mean_spectral_flux	0.000342	56	STD_MFCC	0.000149
24	STD -Standard deviation of period	0.000312	57	Mean_spectral_rolloff	0.000149
25	VAR -Shimmer (apq5)	0.000308	**58**	T_(No)_(CV,CP)	0.00014
26	Mean_F2(Hz)	0.000305	59	VAR_Pitch	0.000131
27	VAR -Degree of voice breaks	0.000296	60	STD_B2(Hz)	0.000126
28	VAR_SSC	0.000288	**61**	VAR_P	0.000126
29	STD -Mean NHR	0.000277	**62**	No_CV	0.000126
30	STD -Degree of voice breaks	0.000268	63	STD -Mean_HNR	0.000125
31	Mean -Shimmer (local_ dB)	0.00026	**64**	R_mean(AV,PP)_(V P)	0.000125
32	Mean -Fraction of locally unvoiced frames	0.000251	**65**	R_(No)_(CV,CP)	0.000125
33	Mean_B2(Hz)	0.000242	66	Kort_SSC	0.000124
			67	R_(time)_(PV,AV)	0.000124

train set. Likewise in AVEC-2013, the performance of the proposed model also computed using only features from the new developed set, the results are shown in Fig 7.7. The results were better than the baseline models for the test set, the MAE= 7.8, RMSE=10.2, and only better in the development for the audio model at MAE=7.9 and RMSE=10.1 and these were lower compared to the the video baseline results. This evidence also showing the usefulness of these features in terms of depression evaluation.

In terms of binary classification approach normal vs. depressed, the models accuracies were 82% and 77% for the development and test sets respectively. Similarly, the better results also was with NW task base model with an accuracy of 84% (for the development) and 90% (for

Table 7.9 AVEC-2014 comparison with baseline results.

Method	Modality	Train Set	Test Set	MAE	RMSE	Acc. %	F1_Score	Precision	Recall	Conf. Matrix
Baseline [246]	**Audio**	Train (NW+FF)	Dev. (NW+FF)	**8.93**	**11.52**	N/A	N/A	N/A	N/A	N/A
	Video	Train (NW+FF)	Dev. (NW+FF)	7.57	9.31	N/A	N/A	N/A	N/A	N/A
	Audio-Video	Train (NW+FF)	Dev. (NW+FF)	6.68	8.34	N/A	N/A	N/A	N/A	N/A
	Audio	Train (NW+FF)	Test (NW+FF)	**10.03**	**12.56**	N/A	N/A	N/A	N/A	N/A
	Video	Train (NW+FF)	Test (NW+FF)	8.85	10.85	N/A	N/A	N/A	N/A	N/A
	Audio-Video	Train (NW+FF)	Test (NW+FF)	7.89	9.89	N/A	N/A	N/A	N/A	N/A
Proposed	Audio	Train NW	Dev. NW	6.5	8.7	84	0.84 (0.86/0.82)	0.85 (0.8/0.9)	0.84 (0.92/0.75)	24 2 6 8
	Audio	Train NW	Test NW	5.8	8.2	90	0.9 (0.9/0.8)	0.92 (0.83/1.0)	0.9 (1.0/0.8)	25 0 5 20
	Audio	Train FF	Dev. FF	6.8	9.7	82	0.82 (0.81/0.83)	0.83 (0.9/0.76)	0.82 (0.73/0.92)	19 7 2 22
	Audio	Train FF	Test FF	6.5	9.8	80	0.80 (0.8/0.8)	0.80 (0.8/0.8)	0.80 (0.8/0.8)	20 5 5 20
	Audio	Train (NW+FF)	Dev. (NW+FF)	**7.1**	**9.4**	**82**	0.82 (0.80/0.84)	0.83 (0.88/0.78)	0.82 (0.73/0.90)	43 5 18 34
	Audio	Train (NW+FF)	Test (NW+FF)	**7.1**	**9.1**	**77**	0.77 (0.78/0.76)	0.77 (0.75/0.79)	0.77 (0.80/0.74)	40 10 13 37
	Audio	Train+Dev. (NW)	Test (NW)	6.1	8.6	84	0.8 (0.83/0.86)	0.85 (0.90/0.80)	0.84 (0.76/0.92)	19 6 2 24
	Audio	Train+Dev. (FF)	Test (FF)	6.9	9.7	78	0.77 (0.73/0.81)	0.82 (0.94/0.70)	0.78 (0.60/0.96)	15 10 1 24
	Audio	Train+Dev. (NW+FF)	Test (NW+FF)	**7.26**	**9.85**	**77**	0.77 (0.75/0.79)	0.78 (0.83/0.73)	0.77 (0.68/0.86)	34 16 7 43

Table 7.10 Depression evaluation with AVEC-2014 corpus comparison to the literature.

Author	MAE	RMSE	Accuracy	F1_score	Precision	Recall
Baseline [246] Audio	Dev 8.93 Test 10.03	Dev. 11.52 Test. 12.56	N/A	N/A	N/A	N/A
Video	Dev 7.57 Test 8.85	Dev 9.31 Test 10.85	N/A	N/A	N/A	N/A
Mitra [188]	Dev 5.87	Dev 7.37	N/A	N/A	N/A	N/A
Pampouchidou [189]	N/A	N/A	Dev 66.0	0.72 (weighted)	0.94 (weighted)	0.59 (weighted)
Morales [190]	Dev 7.56	Dev 9.21	N/A	N/A	N/A	N/A
Zhu [197]	Test 7.47	Test 9.55	N/A	N/A	N/A	N/A
Simantirakiet [187] Dev1: Free-from Dev2: Northwind	Dev1 7.2 Dev2 7.6	Dev1 8.9 Dev2 9.6	N/A	N/A	N/A	N/A
Pérez [196] Meta-model Audio-model	Dev 8.99 Dev 9.35	Dev 10.82 Dev 11.9	N/A	N/A	N/A	N/A
Proposed approach	**Dev 7.25 Test 7.1**	**Dev 9.4 Test 9.4**	**Dev 82.0 Test 80.0**	**Dev 0.82 (0.73/0.90) Test 0.80 (0.80/0.80)**	**Dev 0.83 (0.88/0.78) Test 0.80 (0.80/0.80)**	**Dev 0.82 (0.73/0.90 Test .80 (0.80/0.80)**

the test) compared to 82%(development) and 80%(test) for the FF task model. However, there were no baseline classification scores were reported.

7.2.3.3 AVEC-2016

In the AVEC-2016 the baseline results were provided in terms of MAE and RMSE for predicting PHQ scores and F1-score, precision and recall for the classification task between healthy and

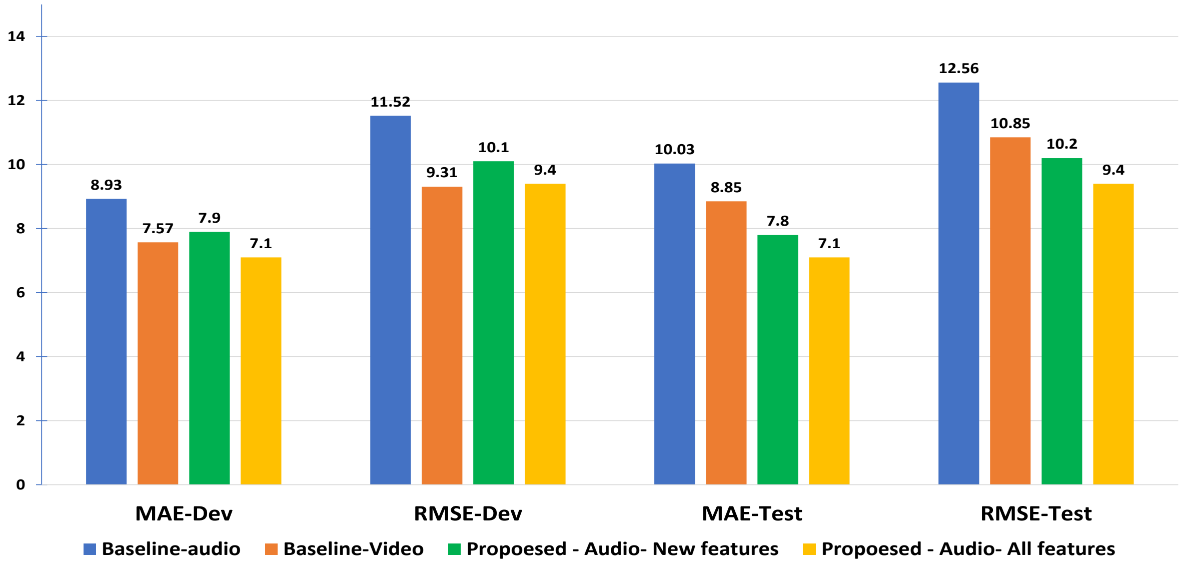


Figure 7.7 AVEC-2014 results comparison with model built using only the new features

depressed patients. Likewise in AVEC-2013 and AVEC-2014, similar approach was adopted for taring and testing a model with the three partitions. Table 7.11 presents the top ranked features as selected by the feature importance process. With this dataset the two proposed models (regression and classification) were superior to the baseline results. Tables 7.12 list PHQ prediction performance, 7.13 shows the classification scores and Table 7.14 compares the results with baseline.

The baseline for audio model that predict PHQ scores achieved (MAE, RMSE) of 5.36, 6.74 and 5.72, 7.78 for the development and test sets respectively, while the proposed model scores 3.6, 5.0 and 4.0, 5.3 which means lower error in predicting PHQ for both development and test parts. These results were also better than the baseline video (development =5.88, 7.13 and test=6.12, 6.97) and ensemble (development=5.52, 6.62 and test=5.66, 7.05) modalities.

Using only features from the new developed set, the results are shown in Fig 7.9. The results were also better than both of the baseline models, the MAE= 4.0, RMSE=5.2 for the development, and MAE=4.3 and RMSE= 5.6 for the test set, and as reported earlier these features plays a major role in improving the performance of the proposed depression evaluation system. The classification performance were also higher than the baseline and for all modalities. The recall metric for the baseline audio were (development, test) (0.85, 0.88), video = (0.42, 0.77) and ensemble = (0.42, 0.77) compared to the proposed audio model obtained recall of 0.91 for

Table 7.11 AVEC-2016 ranked features

Rank	Significant features	Weight	Rank	Significant features	Weight
1	R_VAR_PP	0.013023	37	Mean -Standard deviation of period	0.00461
2	Kort_FBANK	0.009923	38	Mean -Mean period	0.004605
3	Skew_FBANK	0.008672	39	VAR -Jitter (ppq5)	0.004605
4	Mean -Number of voice breaks	0.0085	40	Max_SSC	0.004469
5	VAR -Number of voice breaks	0.007001	41	VAR -Maxmimum pitch	0.00446
6	R_time_(AP,PP)	0.006959	42	STD_energy_entropy	0.004448
7	R_no_(AP,PP)	0.00633	43	VAR_energy_entropy	0.004396
8	R_(mean,no)_PP	0.006214	44	STD -Number of pulses	0.004395
9	VAR -Number of pulses	0.006058	**45**	T_time_PV	0.004372
10	R_no_(PV,AV)	0.006033	46	VAR_F1(Hz)	0.004371
11	R_(T_dif,No_dif)_(AV,PP)	0.005932	47	STD -Shimmer (dda)	0.004346
12	Mean -Mean HNR	0.005673	48	STD -Jitter (ddp)	0.004345
13	VAR -Number of periods	0.005667	**49**	R_(time,No)_(AV,PP)	0.004337
14	Mean -Minimum pitch	0.005661	50	VAR_B2(Hz)	0.004282
15	Kort_MFCC	0.005621	**51**	T_time_(CV,CP)	0.004236
16	STD -Number of voice breaks	0.005594	52	VAR_SSC	0.00422
17	STD -Median pitch	0.005535	53	Mean -Number of periods	0.004184
18	STD -Minimum pitch	0.005422	**54**	R_no_PP	0.004166
19	VAR -Jitter (ddp)	0.005353	55	STD -Mean_HNR	0.004155
20	VAR -Median pitch	0.005345	56	STD -Jitter (rap)	0.004134
21	VAR -Minimum pitch	0.005283	57	VAR -Mean_HNR	0.004134
22	Mean -Number of pulses	0.005258	**58**	R_No(AP,AV)_(VP)	0.00409
23	Mean -Mean NHR	0.005257	**59**	STD_dif_AV	0.004058
24	VAR -Jitter (local)	0.005248	60	Mean_SSC	0.003942
25	R_Mean(AP,AV)_(VP)	0.005218	**61**	P_maxR_T(VP)	0.003905
26	STD -Number of periods	0.005172	62	Mean -Mean pitch	0.00387
27	Min_MFCC	0.005079	63	STD -Shimmer (apq3)	0.003852
28	Mean_F1(Hz)	0.004979	64	VAR -Jitter (rap)	0.003842
29	STD_F1(Hz)	0.004951	65	VAR -Shimmer (dda)	0.003836
30	STD -Maxmimum pitch	0.004913	**66**	No_dif_(AV,PP)	0.003794
31	Mean_Pitch	0.00487	67	STD turn length	0.003777
32	VAR turn length	0.004808	**68**	R_VAR_CP	0.003766
33	\Mean_spectral_spread	0.004728	69	Mean -Jitter (local_absolute)	0.003735
34	Mean -Degree of voice breaks	0.004699	70	Skew_MFCC	0.003733
35	No_CV	0.004656	**71**	R_(T_dif,R_No)_(CV,CP)	0.00373
36	R_VAR_AV	0.004639	72	VAR_B1(Hz)	0.00371

the development and 0.87 for the test set. The classification accuracy were also promising at 90.9% for development and 87.2% for the test. Models developed with this dataset have better performances compared to the models built using the AVEC-2013 and AVEC-2014, this may due to the nature of the task which in AVEC-2016 was more demanding and required additional interactions (semi-structured conversation) compared to the normal (AVEC-2013) and shorter tasks(AVEC-2014).

Table 7.12 DAIC-WOZ PHQ prediction performance.

Training set	**Test set**	**MAE**	**RMSE**
Training	Development	3.6	5.0
Training	Test	4.0	5.3
Training + Development	Test	4.1	5.2

Table 7.13 DAIC-WOZ depression classification performance.

Training set	**Testing set**	**Accuracy (%)**	**F1_Score**	**Precision**	**Recall**	**Confusion matrix**	
Training	Development	90.91	0.90 (0.94 /0.84)	0.92 (0.88 /1.00)	0.91 (1.00 /0.73)	22 3	0 8
Training	Test	87.2	0.87 (0.91 /0.75)	0.88 (0.86 /0.90)	0.87 (0.97 /0.64)	32 5	1 9
Training + Development	Test	85.1	0.84 (0.90/0.70)	0.86 (0.84 /0.89)	0.85 (0.97 /0.57)	32 6	1 8

Table 7.14 Depression evaluation with DAIC-WOZ corpus comparison to the literature.

Author	**MAE**	**RMSE**	**Accuracy**	**F1_score**	**Precision**	**Recall**
Base line [247] Audio	Dev 5.36 Test 5.72	Dev. 6.74 Test. 7.78	N/A	Dev 0.46 (0.68 normal) Test 0.41 (0.58)	Dev 0.31 (0.93 normal) Test 0.26 (0.94)	Dev 0.85 (0.54 normal) Test 0.88 (0.42)
Video	Dev 5.88 Test 6.12	Dev 7.13 Test 6.97	N/A	Dev 0.50 (0.86 normal) Test 0.58 (0.85)	Dev 0.60 (0.86 normal) Test 0.46 (0.93)	Dev 0.42 (0.92 normal) Test 0.77 (0.79)
Ensemble	Dev 5.52 Test 5.66	Dev 6.62 Test 7.05	N/A	Dev 0.50 (0.89 normal) Test 0.58 (0.85)	Dev 0.60 (0.86 normal) Test 0.46 (0.93)	Dev 0.42 (0.92 normal) Test 0.77 (0.79)
Yang [198]	Dev 3.98 Test 5.16	Dev. 4.65 Test. 5.97	N/A	N/A	N/A	N/A
Al Hanai [200] test only with development	Dev 4.97 Dev 5.10	Dev 6.27 Dev 6.37	N/A	Dev 0.43 Dev 0.77	Dev 0.43 Dev 0.71	Dev 0.43 Dev 0.83
Ma [201] development set only	N/A	N/A	N/A	0.52 (depressed) 0.70 (normal)	0.35 1.00	1.00 0.54
Williamson [199] development set only	4.18	5.31		mean (0.81)	N/A	N/A
Proposed approach	**Dev 3.6** **Test 4.0**	**Dev 5.0** **Test 5.3**	**Dev 90.91** **Test 87.2**	**Dev 0.90 (0.94/0.84)** **Test 0.87 (0.91/0.75)**	**Dev 0.92 (0.88/1.00)** **Test 0.88 (0.86/0.90)**	**Dev 0.91 (1.00/0.73)** **Test 0.87 (0.97/0.64)**

7.2.3.4 Language agnostic depression evaluation

To investigate the feasibility of developing a cross-language depression screening tool, and by using the previously proposed models, two scenarios will be examined. The first scenario "SCE-1" will merge all training and development partitions from AVEC-2013, AVEC-2014 and AVEC-2016 into one training part noted as "COM-TR". Then COM-TR is used for building GBC, and GBR models and these models will be evaluated using the three left test parts. The second scenario "SCE-2" will combine all the test sets from all three datasets into one testing part "COM-TS" and used "COM-TR" as a training test. Fig. 7.9 shows the results for both scenarios compared to the previous models in estimating depression severity. There were im-

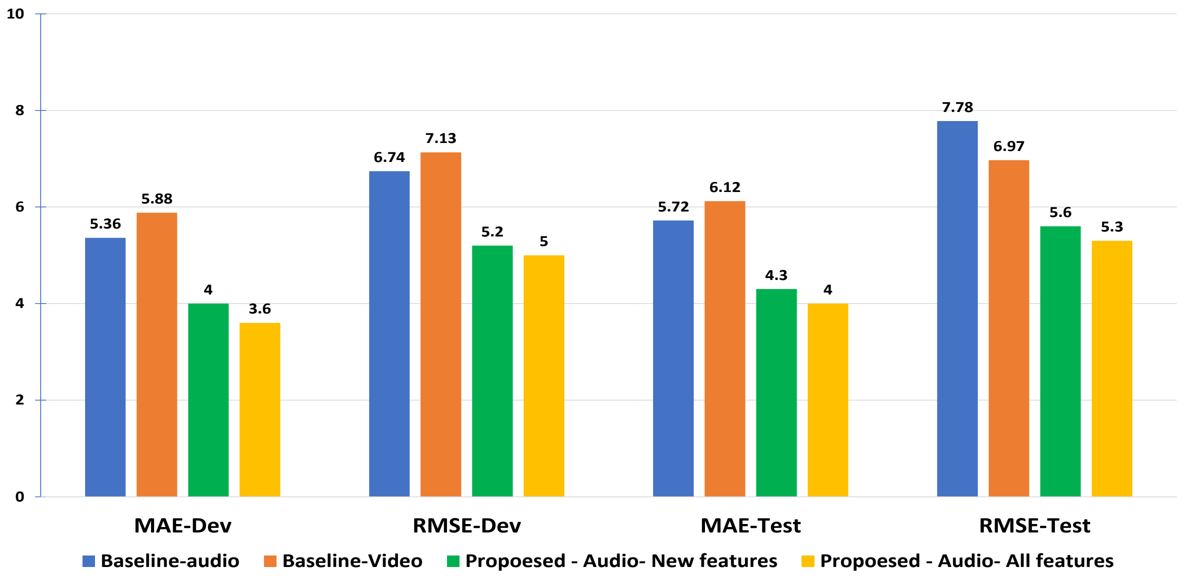


Figure 7.8 AVEC-2016 results comparison with model built using only the new features

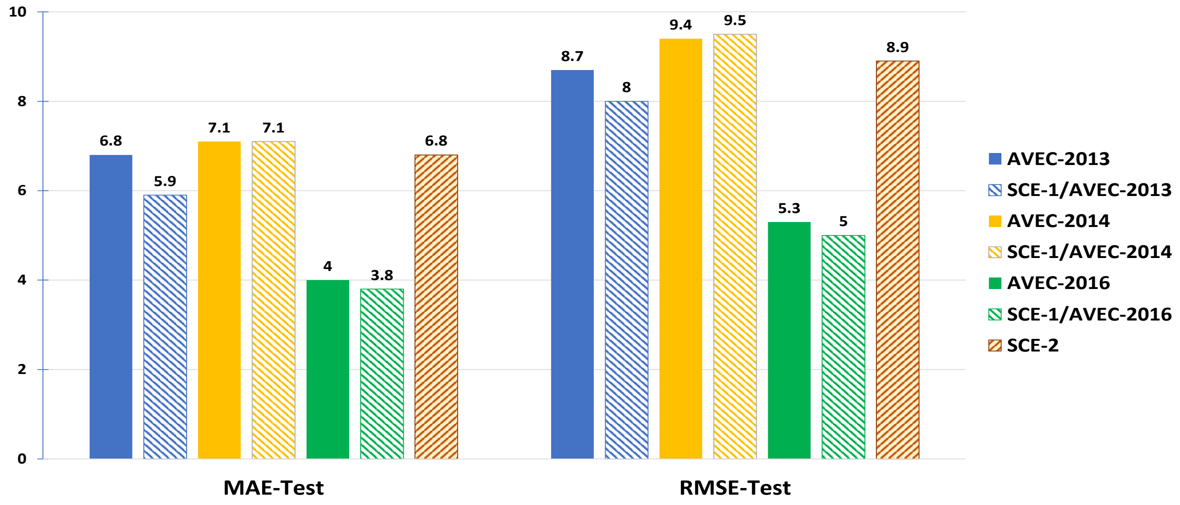


Figure 7.9 All proposed models GBR performances

provements in computing the (MAE, RMSE) for the SCE-1, they both optimised to (5.9, 8) and (3.8, 5) compared to (6.8, 8.7) and (4, 5.3) from the previous AVEC-2013 and AVEC-2016 models respectively. However, there were no improvements in the AVEC-2014, as the MAE remained the same at 7.1, and RMSE was slightly worsened at 9.5 in SCE-1 compared to 9.4. While MAE and RMSE for SCE-2 were 6.9 and 8.9 respectively.

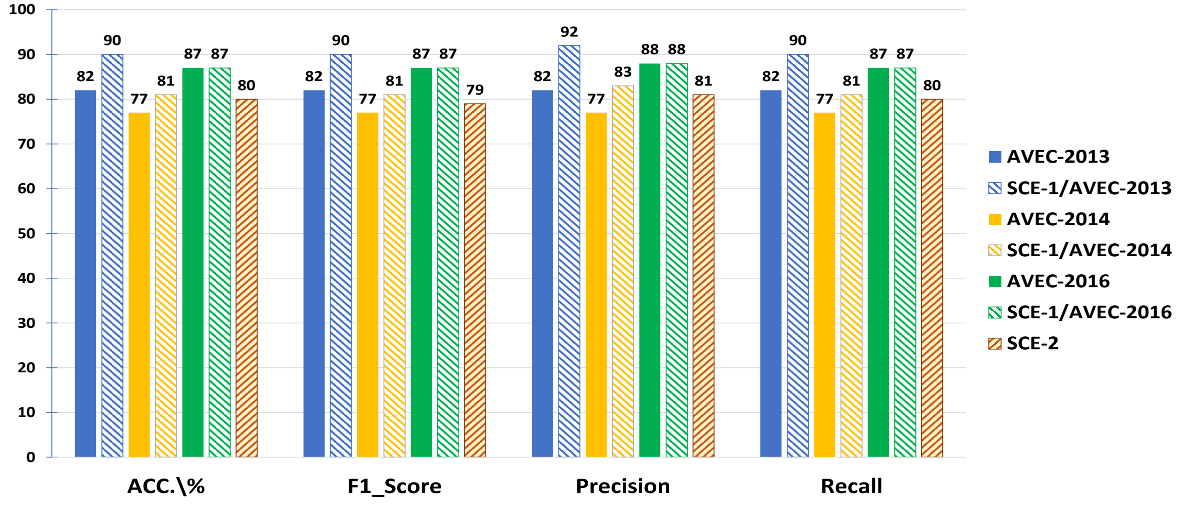


Figure 7.10 All proposed models GBC performances

In terms of GBC performance, in SCE-1 Fig 7.10 shows the classification performances for all the proposed models. This time the SCE-1 depression classification accuracies were 90%, 81% and 87% compared to 82%, 77% and 87% for the AVEC-2013, AVEC-2014 and AVEC-2016 respectively. Only for AVEC-2016 dataset, the accuracy were similar compared to the previous model. While in SCE-2 the accuracy was 80%.

7.3 Discussion

The results from these experiments were promising and support the hypothesis of using only acoustic features to identify depression and estimate and its severity. The newly derived features improved the results as they appear to be of significant values in capturing speech behaviours associated with depression. Table 7.6 list the 56 most important features for AVEC-2013 corpus. These features were selected by the Gradient boosting algorithm using the built-in feature importance, this process performed during the model training with the train set. In this list, there were 20 features from the newly developed variables, and as expected earlier, four of which had overlapped with the features from Table 7.3 of statistically significant features.

Comparing the results to other studies from the literature (listed in Table 7.15). The proposed model outperformed the study conducted by Meng *et al.* [180], the authors used audio, and ensemble approaches and their best results (MAE, RMSE) were 6.94, 8.54 and 8.72, 10.96

Table 7.15 Proposed models performances compared to the literature.

Study	**Dataset**	**Dev. set**		**Test set**	
		MAE	**RMSE**	**MAE**	**RMSE**
Meng et al. [180]	AVEC2013	6.94	8.54	8.72	10.96
Kaya et al. [184]		n/a	n/a	7.84	10.22
Zhu et al. [197]		n/a	n/a	7.58	9.82
Cummins et al. [182]		n/a	10.44	n/a	10.17
Kachele et al. [186]		7.03	8.82	8.72	10.96
Williamson et al. [179]		n/a	n/a	5.57	7.4
Proposed1		**6.6**	**8.9**	**6.8**	**8.7**
Morales et al. [190]	AVEC2014	7.56	9.21	n/a	n/a
Zhu et al. [197]		7.47	9.55	n/a	n/a
Simantirkai et al. [187]		7.2	8.9	n/a	n/a
Perez et al. [196]		9.35	11.9	n/a	n/a
Mitra et al. [188]		5.87	7.37	n/a	n/a
Proposed2		**7.25**	**9.4**	**7.1**	**9.4**
Yang et al. [198]	AVEC2016	3.98	4.65	5.16	5,97
Al Hanai et al. [200]		4.97	6.27	n/a	n/a
Williamson et al. [199]		4.18	5.31	n/a	n/a
Proposed3		**3.6**	**5.0**	**4.0**	**5.3**

for the development and test set respectively. Although the best scores were obtained with a complex approach that fused both audio and video features, yet the proposed model achieved better performances using a simpler model. The proposed model was also better compared to both Kaya *et al.* [184] and Zhu *et al.* [197] studies, despite of the large audio-video feature set used by [184] they reported MAE=7.84 and RMSE=10.22 for the test only, whereas [197] used video-based model and achieved MAE=7.58 and RMSE=9.82 and the results reported for the test set only. These two studies did not report results on the development set; thus, it is difficult to conclude if their model's performances were stables on both sets.

Cummins *et al.* [182] used ensemble audio-video approach to perform depression evaluation on AVEC-2013. They used only RMSE as evaluation metric, and they reported RMSE of 10.44 and 10.17 for development and test sets respectively, which later were improved to 7.4 and 9.49 respectively [183], only this time they used audio-based model. Even with these improvements, the proposed model still managed to provide better results. Other study conducted by Kachele *et al.* [186], they evaluated several approaches using separate audio, video and ensemble approaches to test their models. The best result on the development set was with the video model with (MAE, RMSE) scores of 7.03, 8.82, while the ensemble approach achieved best scores on the test with MAE=8.72 and RMSE=10.96. Despite their complex MLP design algorithm and their adoption of Kalman filter to improve the SVR performance, the proposed system reported better $MAE = 6.6$, $RMSE = 8.9$ for the development set and $MAE = 6.8$, $RMSE = 8.7$ for test set, and these were further improved for the test set to $MAE = 6.2$ and $RMSE = 8.4$ when train the model using both training and development samples. However, the only study that had better results than the proposed model was the study conducted by Williamson *et al.*[179], they reported MAE= 5.75 and RMSE= 7.4 on the test set only. Their audio model was built using parts from the reading task only. These parts were clipped and used to extract the acoustic features. They also removed longer pauses > 0.75 seconds from the clipped segments aiming to reduce the variability of low-frequency dynamics in the formants and delta-MFCC feature extraction process. This method is evaluated using the test set only, also removing longer pauses with > 0.75 seconds may not be generalised to other datasets or other tasks than the reading.

comparing the AVEC-2014 results to others, the proposed model performed better most of the time see Table 7.10. Comparing to Pampouchidou *et al.* [189] study, they conducted binary classification, they reported results with development set only. The accuracy, and the weighted metrics for F1_score, precision and recall were 66%, 0.72, 0.94 and 0.59 respectively, whereas the developed model obtained much better results (in similar order) 82%, 0.82, 0.83 and 0.82 for the development set, and 80%, 0.80, 0.80 and 0.80 for the test set, with no large difference in the results of the two sets, which means that the proposed system is stable and more likely will be generalised to other dataset. The studies from Morales *et al.*[190] and Zhu *et al.* [197] performed single set test, as the first reported MAE=7.56 and RMSE= 9.21 for the development set only, whereas the latter reported MAE = 7.47 and RMSE =9.55. Other study reported on the development set but evaluated each task separately, Simantiraki *et al.* [187] achieved MAE= 7.2, RMSE= 8.9 for "FF" task and MAE=7.6, RMSE=9.6 for "NW" compared to the proposed model with same tasks, this model achieved better results, for FF the model obtained (MAE, RMSE) 6.8, 9.7 and 6.5, 8.7 for the "NW" task. Further study by Pérez *et al.* [196] reported best results with ensemble approach, combining both the audio and video features. The development set results were MAE= 8.99 and RMSE= 10.82 compared to audio model of MAE=9.35 and RMSE= 11.9. The proposed model achieved (MAE, RMSE) of 7.25, 9.4 and 7.1, 9.4 for the development and test set respectively which were better than the previous studies, however, the study by Mitra et al. [188] reported MAE= 5.87 and RMSE= 7.37 for the development set only thus there results with test set were unknown.

The last comparison with literature is with the proposed model developed using the AVEC-2016 corpus see Table 7.14. Yang *et al.* [198] reported respectable results with AVEC-2016, and reported MAE= 3.98, RMSE = 4.65 for the development and MAE= 5.16, RMSE= 5.97 for the test, however, their system far more complex than the proposed model. Al Hanai *et al.* [200] combined several ensemble approaches, the highest MAE and RMSE were 4.97 and 6.27 respectively, while the best classification scores were 0.77, 0.71, 0.83 as an weighted average for F1_score, precision and recall respectively. Their approach had fluctuated in performances between the best in predicting depression severity and best in discriminating depressed vs no normal subjects. Therefore the optimum model is not achieved yet. Furthermore, these results only evaluated with the development set, so it is still unclear how the system will perform if tested with the test set. More study by Ma *et al.* [201] which performed binary classification with the development set only, they reported the results using F1_score, precision and recall for each class (depressed,normal). They achieved results in similar order (0.52,0.70), (0.35,1.00) and (1.00,0.54), although they were able to correctly classify all depressed subjects, they misclassified almost half of the normal group which have the majority in the class distributions (depressed = 11 patients and 22 normal). The last study compared to the proposed model was conducted by Williamson *et al.*[199] who used wide range of different features type including, linguistics, acoustic and facial features. Their ensemble model achieved MAE=4.18, RMSE= 5.31, and an average F1_score of 0.81 for the development set. In this comparison the proposed model obtained higher scores in terms of the predicting PHQ scores or in the binary classification task. The model achieved MAE, RMSE, accuracy, f_score, precision and recall of 3.6, 5.0, 90.91%, 0.90, 0.92, 0.91 for the development set and 4.0, 5.3, 87%, 0.87, 0.88, and 0.87 for the test set in similar order.

7.4 Summary

This chapter demonstrated that acoustic features could be used to detect depression symptoms and estimate their severity. Using the three common depression datasets spoken in two different languages German and English with various tasks and recording environments, the results using the proposed models were promising and outperformed the baseline and most of the studies reported in the literature search. The novel features were developed based on temporal pause and voice segments. The significance of these features were tested, and found that some of them have a statistical significance in the mean difference between the normal group and depressed patients. Several elements from the new set, also selected because of their importance in building and evaluating the model. The performance using the new features was estimated, and they were informative in predicting BDI and PHQ scores with better MAE and RMSE than the baseline scores in the three datasets. Further, the performed task was found to have a strong influence on the model performance, for example, the "NW" from the AVEC-2014 have better discriminate accuracy with MAE= 5.8 and accuracy of 90% than the "FF" with MAE= 6.5 and accuracy of 80%. Finally, the proposed language-agnostic model further improve the results in both regression and classification tasks, this means that the newly developed features have the potential to be used as an objective depression evaluation tool, and these features can be extracted at low cost platforms.

Chapter 8

Automatic screening system for bipolar disorder

8.1 Introduction

Bipolar disease is one type of psychiatric mood disorders, and globally affecting 60 million people according to the World Health Organisation (WHO) [13]. Patients with bipolar disorder frequently experience episodes of depression and mania also undergo transitions to a normal state within these episodes. In a manic state, patients become easily irritated, speaking loudly, experience decreased sleep, and also become hyperactive [257].

The early and accurate diagnosis of the disease means early access to the treatment, which leads to an improved quality of life for bipolar patients. The standard tools for screening bipolar suffer from subjective bias as they rely on a self-evaluation process and on clinicians who interpret the observed symptoms and the patient's responses to the test questions. Thus automatic bipolar screening tools performed on objective measures will be of great benefits to the patients, as this disorder is persistent and requires continuous treatment and monitoring.

The work in this chapter proposes an audio-based approach for detecting the severity of bipolar states. The efficacy of the proposed system is evaluated using the AVEC-2018 bipolar corpus, which is the only accessible dataset that contains audio and video recording for structured interviews of patients suffering from bipolar disease. The rest of the chapter is organised as follows: First the AVEC-2018 bipolar corpus is described and details are presented about the

Table 8.1 Bipolar-AVEC 2018 dataset demographic information.

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

patients, demographic information, performed tasks and the recording procedure. Next is the methodology section, which introduces the proposed system pipeline followed by a statistical analysis applied to the new proposed features to see whether they possess any significance level which can be used to infer bipolar symptomology. After this section, the results were presented and followed by the discussion section; the results section shows how the proposed system performed using the development and test sets samples. While in the discussion, the unbalance classes issue is investigated and addressed. The results are also compared to other studies from the literature.

8.2 Dataset

The dataset consists of short video clips for a semi structured interviews [258]. A total of six recording sessions were administered for most of the subjects and the last session made after the discharge on the third month of the hospitalisation. In each recording session, the subjects performed a number of tasks, for example, describing happy and sad memories, counting up to thirty, express the reason behind visiting the hospital. The Young Mania Rating Scale [259] was used to label the recordings into three groups; Remission, Hypo-mania and Mania. Table 8.1 show the demographic information for the dataset used in this experiment. A total of 218 video and audio files were available, the files were organised into three groups namely: training, development and testing. The owner of the dataset kept the labels for the testing group hidden from the public, so only the training with 104 recordings and the development group with 60 recording were used, see Table 8.2 for more details about the recording sessions and distribution of subjects between the classes.

Table 8.2 Bipolar-AVEC 2018 dataset sessions and classes informations.

Partition		**No. of recordings**	**No. of subjects**	**Average session in seconds**
Training		**104**	**22**	**243.5**
1	Class (1)	25	17	155.71
	Class (2)	38	19	264.79
	Class (3)	41	20	277.3
Development		**60**	**12**	**178.74**
2	Class (1)	(1)	10	153.46
	Class (2)	(2)	9	176.12
	Class (3)	(3)	11	202.94

Legends:
Class(1): Remission: YMRS ≤ 7.
Class(2): Hypo-mania: 7 < YMRS ≤ 20.
Class(3): Mania: YMRS > 20.

8.3 Methodology

The same proposed system pipeline described in chapter 7 is adopted here, only this time, the task is more challenging. The proposed audio system will test whether a multi-class classification task is achievable regarding bipolar states identifications. Fig 8.1 shows the design pipeline for the proposed system used in this chapter.

8.3.1 Feature extraction

The effects of bipolar disorder and its severity on the acoustic features had already been reported in several studies. For example, the disease causes an increase in the median of f0 [260] and the average of F1 and F2 [261]. Further, it is reported that in the depressive state, a higher number of longer pauses were noticed than in hypo-mania conditions [260]. The speech pauses increases with the increase in the severity of the symptoms [262]. These changes are also clinically-known, and evaluated by the psychomotor retardation part in the Hamilton rating scale [91] or YMRS scale [259] as in speech rate evaluation part.

The spectral features were also affected, for example, MFCC coefficients, spectral centroid, spectrum spectral energy, spectral roll-off, spectral flux, slop and entropy [199, 263]. Therefore and based on this evidence, the same feature set that was proposed in chapter 7 see Table 8.3 will be used, and also, the newly developed set of features will be investigated to explore their

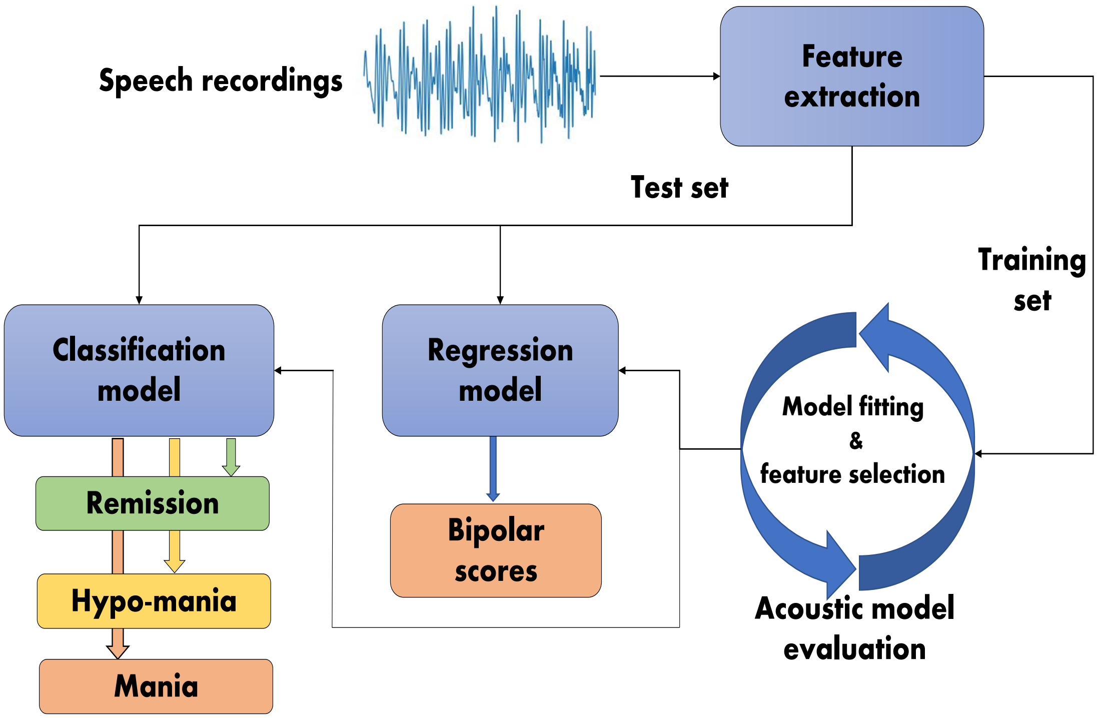


Figure 8.1 Proposed system pipeline for detecting bipolar states

usefulness in dealing with bipolar states classification and YMRS scores prediction.

8.4 Statistical analysis

For this dataset and since there are three compared groups, the ANOVA statistical test was used. Also, the Bonferroni correction test was utilised to resolve for an inflated probability of Type I error (false positive, i.e. rejecting the null hypothesis when it is actually true). Table 8.4 lists the most significant features according to the ANOVA test and only for features that pass the Bonferroni post-hoc test. This means only the features that have a significant mean difference between the three groups were listed, in this way, no feature will be reported even if it has a significant mean difference between just two groups.

The ANOVA test result showed that there are thirteen features with significant mean difference and seven of which were from the newly derived features. Fig 8.2 shows the distributions of these new significant features. The "Mean_CV" is the average of all speech segments time that were above the mean of all pause segments time. The average variables were 3.06, 2.55 and 2.16 for class1,2 and 3, respectively. This indicates that as bipolar state become more severe,

Table 8.3 Final set of acoustic features.

Features	Functions	No. of features
Pitch	Mean,var,std	3
Harmonic-to-noise ratio	Mean,var,std	3
Noise-to-harmonic ratio	Mean,var,std	3
Shimmer (six scales)	Mean,var,std	18
Jitter (five scales)	Mean,var,std	15
Number of voice breaks	Mean,var,std	3
Fractions of locally unvoiced frames	Mean,var,std	3
Degree of voice breaks	Mean,var,std	3
Number of voice breaks	Mean,var,std	3
Number of periods	Mean,var,std	3
Auto_correlation	Mean,var,std	3
Number of pulses	Mean,var,std	3
MFCC	Min,max,skewness,kurtosis,mean,var,std	7
Fbank)	Min,max,skewness,kurtosis,mean,var,std	7
SSC	Min,max,skewness,kurtosis,mean,var,std	7
Intensity	Mean,var,std	3
F1(Hz)	Mean,var,std	3
Intensity at F1	Mean,var,std	3
B1(Hz) bandwidth_F1	Mean,var,std	3
F2(Hz)	Mean,var,std	3
Intensity at F2	Mean,var,std	3
B2(Hz) bandwidth_F2	Mean,var,std	3
ZCR	Mean,var,std	3
Energy	Mean,var,std	3
Energy_entropy	Mean,var,std	3
Spectral_centroid	Mean,var,std	3
Spectral_spread	Mean,var,std	3
Spectral_flux	Mean,var,std	3
Spectral_rolloff	Mean,var,std	3
Spectral_entropy	Mean,var,std	3
New developed features	Mean,var,std, ratio, difference	99
Total		228

the speech of active voices will be reduced. The second feature is the "R_{mean}_CP" which also was significant in the AVEC-2014 statistical test. This evidence shows that between different datasets and different spoken languages this newly developed feature associated with the severity of the disease, similar to AVEC-2014, the large value here suggests lower severity bipolar condition. The mean for the three groups were 1.2, 0.68 and 0.58 for class 1,2 and 3 respectively. Next is "R_{mean}_AV" which is the ratio of the average time for voice segments to the average time of all active voices. A larger value means less active voice segments were pro-

Table 8.4 AVEC-2018 ANOVA statistical significance test.

Rank	**Features**	**DF**	**F-test**	**P-Value**
1	$Mean\_CV$	101	10.481	0.000073
2	STD_F1Hz	101	9.617	0.000
3	VAR_F1Hz	101	8.130	0.001
4	Max_FBANK	101	7.934	0.001
5	Mean_FBANK	101	7.745	0.001
6	R_{mean}_CP	101	7.005	0.001417
7	R_{mean}_AV	101	5.652	0.005
8	R_{no}_AV	101	5.514	0.005
9	Min_FBANK	101	5.456	0.006
10	$Mean\_AV$	101	5.313	0.006
11	VAR-number-of-voice-breaks	101	5.184	0.007
12	$P_{max}R\_T(VP)$	101	5.048	0.008137
13	$R_{no}\_(PV, AV)$	101	4.617	0.012056

Legends: STD: Standard deviation, VAR: Variance, DF: Degree of freedom,
F-test: F-statistic score for ANOVA, class(1) n=25, class(2) n =38, Class(3) n=41.

duced by the individual, and interestingly this was the value for Mani group with the mean of 0.6 and slightly lower at 0.59 for Hypo-mania patients while lowest at 0.54 for the Remission group.

The fourth important variable was "R_{no}_AV" which is the ratio of the number of active voice segments to the total voice segments time. According to this feature, a larger value indicates severe condition; the mean of the three classes were 0.19, 0.2, and 0.24. The fifth new informative element is "Mean_AV". This measure the average of active voice segments time and larger value may refer to an active speaker, thus a lower bipolar condition. Class0 have mean = 3.4, class 1 = 3.2 and class 3 = 2.6. Further, "P_{max}R_T(VP)" measures the ratio of max pause segment time to the total of all voice and pause segments time. The mean value were 0.054, 0.035 and 0.032 for class 1,2 and 3 respectively. The last significant feature was "R_{no}_(PV,AV)" which is the ratio of the total number of passive voice segments to the total number of active voice segments. This behaviour also confirms that in sever bipolar states, this variable will have a lower value. This feature had mean values of 2.15, 1.77 and 1.72 for class 1, 2 and 3 respectively.

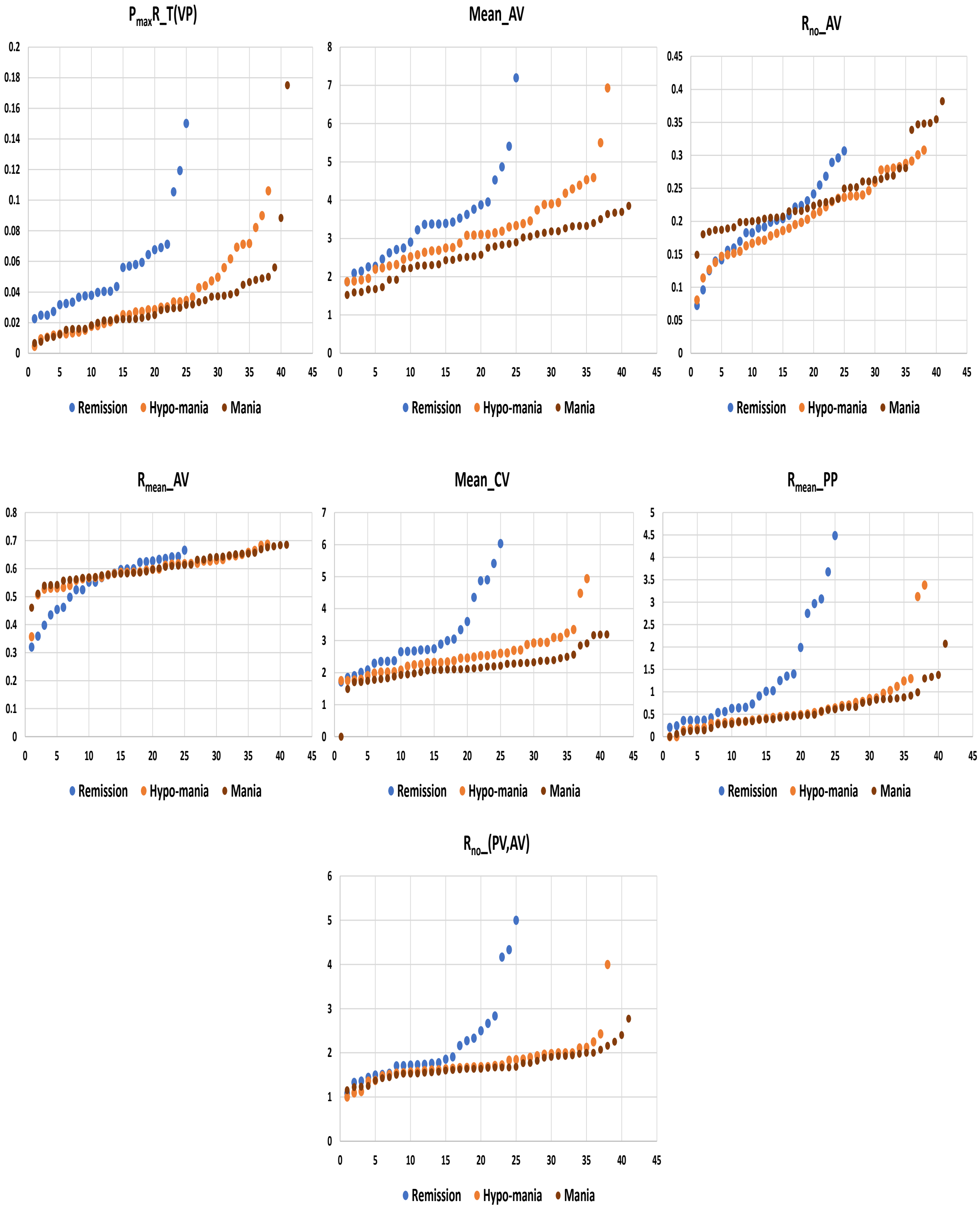


Figure 8.2 AVEC-2018 the distributions of the new most significance features based on the t-test.

8.5 Results

The task for the AVEC-2018 challenge was to predict the three states of the bipolar condition, i.e. remission, hypo-mania and mania. The owner of the dataset kept the labels for the test set hidden from the public. Each team who participated in the challenge will have to send the test set class prediction to the AVEC-2018 team [258], and they will send back the results in terms of the unweighted average recall (UAR) Only five attempts were allowed for each team.

The Gradient Boosting Classifier (GBC) and Gradient Boosting Regression (GBR) algorithms were used to evaluate the proposed acoustic system for detecting bipolar conditions. The model trained using the training partition, and the feature importance technique was used to define which features are crucial during the training phase. Table 8.5 shows the list of 67 most significant features ranked by their importance weight. There were 36 of which (displayed in bold) were from the new extracted feature set, which points out the value of these features in capturing bipolar symptoms. Further, five from these new variables already showed significant value as per the ANOVA statistical test Table 8.4, another evidence which highlights the importance of these features.

The results are listed in Table 8.6. In the first attempt "Propose−1", the UAR for the development set was far better than all baseline modality results, the model achieved 71.6% compared to 55.03% for audio, 55.2% video and 63.49% for the ensemble approach. However, the proposed model performed poorly with the test set and had UAR of 42.59% compared to 50%, 46.3% and 57.41% for the audio, video and ensemble, respectively. This fluctuation in model performance due to over-fitting on the development partition. Also, another challenge was the unbalance class distribution for the samples in the training set (Class 0 =25, class 1=38 and 41 for class 3). In the second attempt and to address the unbalanced training data and to provide a more generalised model can perform well in both partitions. A total of 38 samples randomly were moved from the training to the development set taking into consideration of moving samples from the classes with the majority. The new training set now has 66 samples with a distribution of 21, 24, and 21 for class 1, 2 and 3 respectively, while the development set increased to 98 samples distributed into class1 =22, class 2 = 35 and 41 for class3. Further, the 35 samples that moved were from 8 subjects, and no samples from a unique subject are overlapped between the two sets. This is done to avoid the bias in the model training, which might be reflected on

Table 8.5 AVEC-2018 most significant features

Rank	Significant features	Weight	Rank	Significant features	Weight
1	Kort_SSC	0.009763	**35**	**R(T_{mean}, No_{dif})_(CV,CP)**	0.004585
2	Skew_SSC	0.009072	36	Mean_T_intensity	0.004496
3	Max_SSC	0.008808	**37**	**VAR_CV**	0.004487
4	STD_F1(Hz)	0.008605	**38**	**T$_{time}$_PP**	0.004435
5	Mean -mean-auto-correlation	0.008165	**39**	**R$_{No}$_(AV, PP)**	0.004425
6	Mean_FBANK	0.00791	**40**	**R$_{mean(CV,CP)}$_(VP)**	0.00441
7	**R(mean,no)_AV**	0.007828	**41**	**T$_{time}$_AV**	0.004358
8	**Mean_CV**	0.007715	**42**	**T$_{time}$_(CV,CP)**	0.004332
9	**R$_{Mean(AP,AV)}$_(VP)**	0.007337	**43**	**Mean_V**	0.004315
10	Mean_I1	0.006544	**44**	**R$_{max}$R_T(VP)**	0.004307
11	**T$_{time}$_CV**	0.006542	**45**	**No_AV**	0.0043
12	**T$_{time}$_PV**	0.006452	**46**	**T$_{No}$_(CV,CP)**	0.004287
13	**R(T$_{dif}$, No$_{dif}$)_(CV,CP)**	0.006424	**47**	**T$_{dif}$_(CV,CP)**	0.00423
14	VAR_F1(Hz)	0.006201	48	Mean−Shimmer (dda)	0.004211
15	**R$_($$T_{dif}$, No$_{dif}$)_(AV,PP)**	0.00616	49	STD −Jitter (ddp)	0.004199
16	STD−Jitter (local_absolute)	0.006096	**50**	**STD_AV**	0.004198
17	Min_FBANK	0.006074	**51**	**VAR_AV**	0.004115
18	Mean -mean harmonics-to-noise ratio	0.005888	**52**	**R$_{mean}$_AV**	0.004058
19	STD_I1	0.005875	**53**	**R$_{(time,No)}$_(CV, CP)**	0.004031
20	Max_FBANK	0.005754	54	STD_spectral_spread	0.003937
21	**R$_{T(AP,AV)}$_(VP)**	0.005636	**55**	**R$_{mean(CV,CP)}$_(VP)**	0.003926
22	Min_SSC	0.005628	56	STD -Number of voice breaks	0.003883
23	VAR_I1	0.005566	**57**	**R$_{(T_mean,No_dif)}$_(CV,CP)**	0.00379
24	T$_{mean}$_time_(AV,PP)	0.005521	**58**	**No$_{dif}$_(AV,PP)**	0.003741
25	Mean -Shimmer (local)	0.005422	59	Max_MFCC	0.003733
26	**Mean_AV**	0.005121	60	VAR -Jitter (ddp)	0.003718
27	**R$_{no}$_AV**	0.005084	61	STD -Jitter (local)	0.003704
28	**R$_{Time}$_AV**	0.005063	62	STD -Standard deviation	0.003675
29	**No_CV**	0.004965	**63**	**R$_{mean}$_CP**	0.003658
30	Mean -Shimmer (apq5)	0.004896	64	STD -Minimum pitch	0.003653
31	**Tmean_time_(CV,CP)**	0.004839	**65**	**R$_{(time,No)}$_(AV,PP)**	0.00364
32	**R$_{(R_time,No_dif)}$_(CV,CP)**	0.004717	66	VAR -Jitter (rap)	0.003607
33	Min_MFCC	0.004698	67	Mean_F2(Hz)	0.003601
34	**Mean_AP**	0.004675			

Table 8.6 Bipolar evaluation results compared to the literature.

Author	Model	Partition	% UAR
Baseline [259]	Audio	Development	**55.03**
	Video	Development	55.2
	Ensemble	Development	63.49
	Audio	Test	**50**
	Video	Test	46.3
	Ensemble	Test	57.41
Du [264]	Ensemble	Development	65.1
Xing [267]	Ensemble	Development	86.77
	Ensemble	Test	57.41
Yang [272]	Ensemble	Development	71.41
	Ensemble	Test	57.41
Syed [263]	Ensemble	Test	57.41
Proposed-1	Audio	Development	71.6
	Audio	Test	42.59
Proposed-2	Audio	Development	**66**
	Audio	Test	**53.7**

Table 8.7 Bipolar-AVEC 2018 PHQ prediction performance.

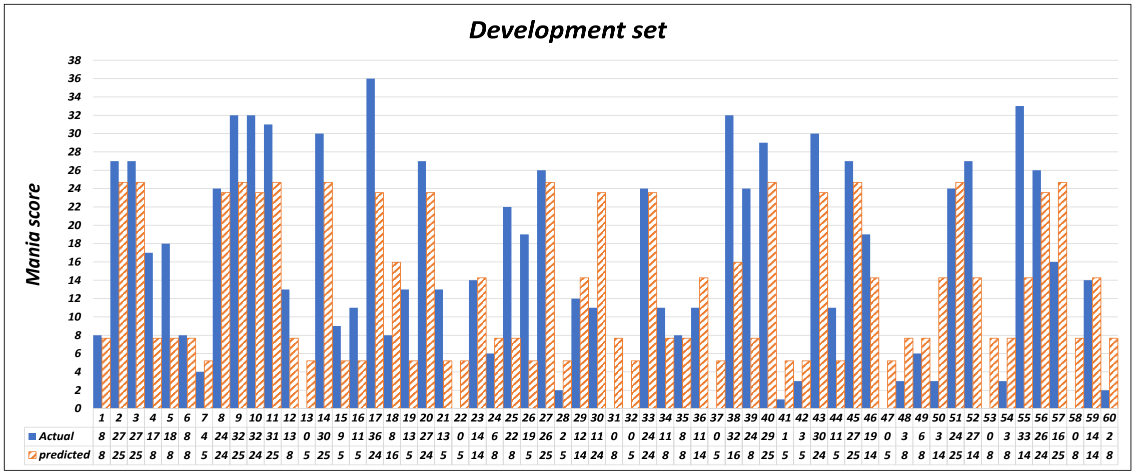
Training set	Test set	MAE	RMSE
Training	Development	5.7	7.2

increasing performance with development set but lead to diminishing the scores on the test set.

Although the results from this scenario "Propose−2" had reduced the UAR for the development set from 71.6% to 66%, it increased the test scores from 42.59% to 53.7%, even with the development decreased performance, it still managed to outperform all development baseline modalities. The improvement on the test set also was better compared to the baseline audio (50%) and video (46.3%) modalities but yet still lower than the ensemble with 57.41%. Additionally, YMRS scores estimation was performed for the development set, and the result are listed in Table 8.7 and in Fig 8.3 which shows the actual vs predicted scores. The proposed model achieved MAE= 5.7 and RMSE= 7.2; however, there were no baseline results for this task, yet these results optimistically respectable results when compared to chapter 7 results, they were better than the AVEC-2013 (MAE= 6.6, RMSE= 8.9), and AVEC-2014 (MAE= 7.1, RMSE= 9.4) but lower than AVEC-2016 (MAE= 3.6, RMSE = 5.0).

8.6 Discussion

This experiment aimed to build an automatic bipolar conditions screening tool using audio features. The results from the ANOVA statistical significance test and the feature importance ranking method showed that the new set of features were informative in this task. The top 36 selected features associated with bipolar symptomology, for example, "Mean_CV", measures the average of cross voice segments, which attempts to describe the speaking activity by measuring the average of all voice segments that were above the means of pauses time; a higher value indicates an active speaking pattern, and in this dataset a highest Mean_CV of 3.0 seconds was for the remission group, a less than 2.5 seconds value for hypo-mania and a lowest Mean_CV at 2.1 was for the mania patients. Another important variable from the new set is the "Mean_AV" which describes the average of active voice segments spoken by the subject. This feature characterises the speaking behaviour, where larger active indicates larger voice segments were produced compared to the mean of the utterance and smaller indicates less activity speaking pattern. For class remission, which is a low bipolar state, the Mean_AV for this class was 3.4 seconds higher than class hypo-mania more severe condition, therefore, reduced active speech at Mean_AV of 3.2 seconds. That even higher compared to class mania, the lowest state



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
■ Actual	8	27	27	17	18	8	4	24	32	32	31	13	0	30	9	11	36	8	13	27	13	0	14	6	22	19	26	2	12	11
▨ predicted	8	25	25	8	8	8	5	24	25	24	25	8	5	25	5	5	24	16	5	24	5	5	14	8	8	5	25	5	14	24

	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
■ Actual	0	0	24	11	8	11	0	32	24	29	1	3	30	11	27	19	0	3	6	3	24	27	0	3	33	26	16	0	14	2
▨ predicted	8	5	24	8	8	14	5	16	8	25	5	5	24	5	25	14	5	8	8	14	25	14	8	8	14	24	25	8	14	8

Figure 8.3 AVEC 2018-Bipolar YMRS scores estimation, the actual vs predicted for the development set

with decreased speech activity at lowest Mean_AV of 2.6 seconds. This implies for the rest of the new significant features. It is also evident that low speech rate, longer pauses and reduced in voice segments size were linked to severe mood disturbances [74, 129]. However, the new features provide more in-depth details to the speaking behaviour than the common functions, for example, a large number of small voice segments would probably have the same average of less number of voice segments with larger voice chunks; thus the average is probably useless in this case. On the other hand, the new feature set can characterise this behaviour in more details, for instance, the "Mean_CV" compare the average of these voice segments to the mean of pauses. The "Mean_Av" which estimate only the mean of the active voice segments, further, the "R_{Time}_AV" computes the ratio of the active voice segments to the total number of all voice segments, and so on for the rest the new variables.

Although the results on the test set were lower than the baseline by a small margin, the proposed model is simpler, based only on audio features compared to the more complex audio-video system. The optimised results with "Propose−2" over "Propose−1" was first due to balancing the samples between the training classes, this means that the GB-CLF algorithm more-likely is constructed in such way to possess an even class prediction probability for all of the three classes and for each attempt. This approach will lead to better performance on the test set, whether it has balance or imbalance classes. Also, this design will eliminate the need for adjusting the class probability threshold towards the class with the majority of samples. Changing the class probability may lead to improve the performance for example on the development set, however, this will not guarantee better performance on a test set, because the adjustments that were made on one set probably will not fit other data. Secondly, the model improved using a larger developed set; thus, a more generalised approach is more likely to perform better on the test set.

How this model compares to other studies is shown in the same Table 8.7. The proposed system had a better result only compared to Du *et al.* [264] study, who applied InceptLSTM, which is a type of a deep neural network approach to identify the three states of bipolar disorder. Their system combined both capabilities of CNN and LSTM and tested on the AVEC-2018 dataset. They introduced another method called the severity-sensitivity loss to improve the cost of reducing the distances for the samples within the same class and increasing the gaps between the classes themselves thus improve the performance of their proposed model . Du et al. evaluated three SVM models, the first one developed using MFCCs and its related features, the second model was trained using eGaMAPS features[265] while the last SVM built using DeepSpectum variables [266]. The reported results indicate that InceptLSTM model is better compared to the three SMV models in terms of detecting the bipolar conditions. The best model achieved a UAR of 65%. This result was generated from the development set, and it is unclear how the model will perform on the test data.

Comparing to Xing *et al.* [267], who adopted an approach inspired by Gone *et al.* [268] system, as the latter developed automatic depression screening system. The system combined several features from audio, video and linguistic. Gone *et al.* reported that the proposed system performed well when ordering the extracted features to match the content of the spoken language. Xing *et al.* used Google Cloud Platform (GCP) to transcribe the AVEC-2018 recordings, later the authors assorted their sets of features into three groups natural, positive and negative based on the valence of the utterances. Xing *et al.* extracted a large set of features including MFCC and eGeMAPS from the audio. From SALAT toolkit [269], they added a set of linguistic features. From the video, they extracted the MHH histogram based on action units, and finally, they used the Faceplusplus toolkit [270] to add Ekman's seven emotions variables. The authors used ANOVA statistical significance as a feature selection procedure. The proposed system was evaluated using eXtreme Gradient Boosting (XGBoost) algorithm [271] achieving outstanding UAR of 86.77% for the development set; however, the performance decreased to 57.41% when predicted the test partition. This sharp drop suggests an overfitting issue; this means that the model constructed with a set of variables such as tuning parameters and a group of features that maximises the performance with development samples. Thus when the model tested with new samples from the test set, these same variables failed to provide a similar result, because of the variation of the samples between the two sets. The proposed model had weak generalisation property; therefore, its reliability is questionable.

Another study by Yang *et al.* [272] introduce arousal features for the task of automatic bipolar severity estimation. They showed that when applying a histogram to arousal features provide useful information for discriminating bipolar states. Due to missing the arousal points in AVEC-2018 dataset, the authors developed LSTM-RNN model using the dataset from the

AVEC-2015 to predict the arousal scores for each of AVEC-2018 recordings by concatenating all scores using histograms to create a unified global representation for the arousal ratings for all samples. The authors extracted several audio features using Opensmile [191], visual elements, including facial action units, and body motion using Openpose toolkit[273]. Due to the large number of extracted features, Yang *et al.* used SVM classifier, correlation-based feature reduction procedure [274] and brute force forward search algorithm [275] to optimise their feature set. The final evaluation model built utilising fusion-based approach for both the random forest classifier and DNN. They reported UAR of 71.4% for the development and also dropped to 57.41% for the test set. Despite using several modalities developed with a number of classifiers, they only manage to match the baseline results.

Finally, the study by Syed *et al.* [263], in their experiment, they proposed features called "turbulence-features" to capture the variations of the fundamental frequency (F0) contour as they hypothesis that these features will be useful in estimating the severity of bipolar states. They used COVAREV toolkit [276] to extract the F0 using different window lengths of 0.5, 1.0 and 2.0 seconds and with overlapping of 0.2, 0.4 and 0.8 seconds respectively. The turbulence features represent the ratio of the maximum absolute value for the signal and its root mean square value. Then they applied statistical functions including 10^{10}, 25^{10}, 50^{10} and 75^{10} and the mean trimmed with 5%. They used the Opensmile toolkit [191] to add number of paralinguistic and prosodic features to their audio modality. They also extracted features aims to capture the emotional and movement changes might occur due to the severity of bipolar states. These variables extracted using OpenFace toolkit [277], including $3-$D head-pose, the vertical and horizontal eye-gaze angel and six facial action units. Finally, they introduced the Greedy Ensembles of Weighted Extreme Learning Machines (GEWELMs) as their evaluation model. GEWELMs developed using an arbitrary number of Weighted Extreme Learning Machines [278] WELM algorithm. Later they selected only the 90^{10} percentile of all WELMs tested with on the development set but were higher than an unspecified threshold. The best UAR result they achieved using the AVEC-2018 test partition was 57.41%. The proposed approach is considered costly in terms of the number and type of features it needs to be able to perform. Even with this level of complexity, they only manage to match the baseline result. Further, they did not report the UAR on the development to see whether their model had suffered from the over-fitting issue or not.

8.7 Summary

In this chapter, a simple audio system for the task of classifying the severity symptoms of bipolar disease was proposed, i.e. to identifies each recording from the AVEC-2018 corpus into one of the three classes: remission, hypo-mania and mania. Also the YMRS scores prediction model was proposed based on audio modality, the results were compared to the PHQ and PDI scores prediction models from the previous chapter.

The developed "speech activity behaviour" features were used in the proposed model, and they have proven to be informative as per the ANOVA statistical test results and in the feature importance selection procedure.

This system was able to outperform the results of the baseline audio and video modalities. The issue of the unbalanced data classes was investigated and its effects on the performance behaviour in both of the development and test sets was illustrated. To this end, the proposed balancing to the dataset in the training partitions by removing samples from the classes with the majority of samples and add them to the development set had yielded better performance with the test. This design means the system is more likely to be generalised to other data.

Chapter 9

Conclusions and future work

9.1 Conclusions

This thesis investigated an audio-based detection of early signs of dementia and mood disorders. To this end, a simple approach solely based on acoustic features had been proposed and showing robustness in identifying dementia of Alzheimer disease, later this system was further developed and had demonstrated the ability of longitudinal monitoring for AD. Another approach was designed to extract useful information from patients conversations, and this aided in discriminating symptoms of functional memory disorder from other patients having neurodegenerative diseases. Furthermore, depression and bipolar disorders audio-based screening proposals were introduced and showed potential for the use as language-agnostic models.

Recent reports had shown that both dementia and mood disorders are causing a considerable rate of disability not only in developed countries but also worldwide. In the UK, dementia is affecting more than 850,000 people with overall care cost exceeding £26.3 billion per year, and the number of patients expected to be more than 2 million by 2051 [8]. While there are 2.69 million people (4.5% of the total population) diagnosed with depressive disorder [14] with an annual cost ranges between £7-£9 billion [15, 16].

Dementia is a group of symptoms in which there is a decline in memory, reasoning, judgement, behaviour and the ability to execute daily activities. Dementia also affects the speech and language performance; although subtle changes may be observed prior to the early stage, however, with disease progression, patients will suffer from difficulties in communication and expressing their needs. As a result, patients will isolate themselves from the surrounding people, and others will develop depression and show aggressive behaviours; eventually, dementia leads to death. Whereas depression is a psychiatric mood disorder caused by a sudden stressful event affecting an individual's life, for example, losing a beloved one, losing jobs etc. This disorder leads to a continuous feeling of sadness, negativity, and makes it hard to handle everyday responsibilities. Depression often triggers suicidal thoughts to end one's life [18, 19]. It has been confirmed that at least 400000 people die every year from committing suicide, and those fulfil the clinical diagnosis of depressive disorders [21, 22].

In regards to dementia, there is no cure; however, several disease modification treatments exist that help in slowing down disease progression. These drugs are effective when consumed as early as possible, and before the irreversible brain damages occur. Diagnosing dementia at the earlier stage is a challenging task, due to the overlapping symptoms that concerning memory issues with similar ones from other disorders, for example, the functional memory disorders, depression, or even from the normal ageing process. The current tools used to identify those patients with a higher risk of developing dementia are either costly; for example, Positron Emission Tomography (PET) or invasive, for example, the CerebroSpinal Fluid (CSF). Further, the cognitive test batteries that currently used for screening patients use the traditional pen and papers, established based on the English language, showing learning effects which limits the number of possible administration, and lack sensitivity or specificity. Similarly, depression diagnosis tools and tests based on interviews assessment, for example, Hamilton Rating Scale for Depression and suicide probability scales, both methods measuring the severity of symptoms and behaviours observed in both conditions thus the patients will obtain a test score range based on the perceived symptoms. Using this method of assessment is not straightforward; the given scores are sensitive to the patient's ability to express their symptoms, moods, and cognitions honestly and willingly. Consequently, collecting diagnostic clues is a time-consuming procedure and involves a significant amount of clinical training, experience, and certification to induce satisfactory results.

As a result, developing a cheap, non-invasive and non-intrusive, automatic, and objective tool that can be used frequently without learning effects, remotely applicable, reliable, and easily administrated is on high demand by healthcare providers. This tool can bring tranquillity for those at low risk of developing dementia, and at the same time to speed up the process of providing the right medications to those who most likely demented. Further, where most of tests founded based on the English language, translating it to other languages may not work very well due to culture differences, level of educations, etc. Therefore language-agnostic screening or testing systems are also importantly needed.

Based on that, this thesis has adapted the latest speech signal processing and machine learning techniques to develop and propose a simple systems for detecting early signs of dementia, monitoring its progression, and screening for depression and bipolar disorders. To this end, the thesis attempted to find answers to number of research questions which were listed in chapter 1.

9.1.1 The feasibility of developing an automatic, low cost and simple system to aid the doctors in identifying early signs of dementia?

The first research question was how feasible it is to develop a simple and automatic system to detect early signs of dementia diseases. To answer this question first is to select what type of model that fulfil the simplicity, robustness and low-cost targets of the research question. The literature search suggests that audio modality is a better fit for the research question requirements than video or the ensemble approach. The video-based approach required good quality hardware for recording video clips for the patients/participants that are needed for the analysis purpose. These devices normally more expensive than audio acquisition hardware besides the audio model is less intrusive in terms of maintaining patient's privacy.

Secondly, chapter 4 introduced the pipeline of an audio-based system that proposed to detect early signs dementia of Alzheimer Disease (AD). The system consist of prepossessing and feature extraction unit, then features were used to train number of well known machine learning classification algorithms. This system developed using the dementiabank dataset, in which participants performed the Cookie Theft Picture description task. Using acoustic features extracted from the short audio recordings, the proposed system achieved promising classification accuracy of 94.7% when classifying between AD patients and healthy control participants (HC), with sensitivity and specificity of 97% and 91% respectively. This finding answered the first research question and was reported in [240].

Finally, chapter 6 described a novel automatic dementia detection system developed with collaboration from neurologists at Royal Hallamshire hospital in Sheffield. This study demonstrated that purely acoustic features, extracted from recordings of patients' answers to a neurologist's questions in a specialist memory clinic can support the initial distinction between patients presenting with cognitive concerns attributable to progressive neurodegenerative disorders (ND) or Functional Memory Disorder (FMD, i.e., subjective memory concerns unassociated with objective cognitive deficits or a risk of progression). The study involved 15 FMD and 15 ND patients where a total of 51 acoustic features were extracted from the recordings. Using three types of feature selection approaches the recursive feature elimination, feature importance and Mann-Whitney u-tests, result in reducing the number of informative features to 22. This system had showed very promising result with an accuracy ranged between 93% and 97% with an average of 96.2% for five different classifiers SVM, Random forest, Adaboost, multilayer perceptron and stochastic gradient descent. Further, another validation scheme was developed to simulate the behaviour of the system when testing with larger dataset, as the data augmentation scenario showed that there was 5% decline in performance when increasing number of samples from 30 to 230 (i.e., more than 750% increment) and for each 100% increment in size the performance decreased by 0.67% which shows that the proposed approach did not deviate badly, and the classification accuracy ranged between 87% and 92%. The findings from the work also been published in [252] and contributed to answer the first and the last research question,

9.1.2 The feasibility of designing low cost and simple system that assists the doctors in predicting the severity of dementia and monitor its progression?

To answer this question first is to define a common test used in measuring the severity of dementia symptoms, for example, MMSE scores. Secondly, a hypothesis needs to be fulfilled, in which if there is a system that is capable of replicating the MMSE scores using audio modality generated from patients suffer from (AD), then this system manifests the assumption and therefore provide the answer to this question. Chapter 5 described a system that showed respectable results in predicting MMSE scores. The system was developed using DementiaBank dataset, with different evaluation scenarios to illustrate the effectiveness of this system not only in predicting the current MMSE scores but also an estimation for future scores. By using three visits from DementiaBank, the proposed system predicted the longitudinal MMSE scores with MAE of 3.1, 2.6, and 3.7 for visit1, 2 and 3 respectively. Other scenarios were tested including predicting a future MMSE for a visit3 using visit2 dataset, and the system achieved MAE of 2.25, while, the last model proposed achieved MAE of 2.18 when combing both visit1 and 2 samples to predict visit3 MMSE scores. In chapter 5 and using the speech recordings from the same dataset, but only this time increased the task complexity. The previous task only classified between AD and HC, and know included participants from Mild cognitive impairment MCI. With this new level of complexity the system maintained outstanding results, as for visit1 comparing between AD vs HC, HC vs MCI, and AD vs MCI the accuracy were 91.2%, 93.7% and 95% respectively, this also was optimised and by using the synthetic minority oversampling technique (SMOT) these results improved to 94%, 96.5% and 96.6% respectively. Although the SMOT method only used to overcome the unbalanced samples between the compared classes, it improved the performance of the proposed system. The finding of this work were published in [241] addressing the second and last research questions.

9.1.3 The possibility of developing an objective tool that can be used to identify depression and estimate its severity?

Chapter 7 illustrated novel features developed based on the temporal voice and pause segments features. The new activity features proven to be effective in capturing speech behaviour associated with depression disorder, both in statistical tests and in the feature selection step when the proposed models were constructed. These approaches were developed using three publicly available depression datasets knows as AVEC-2013, AVEC-2014 and AVEC-2016 audio/video challenges. The first two datasets were spoken in Germany while the last one collected from native English participants. Using only the speech activity features the proposed depression evaluation models outperformed the baseline challenge results with MAE of 7.4 and 7.8 compared to AVEC-2013 of 8.66 and 10.35 for the development and test sets respectively, and proposed MAE of 7.9 and 7.8 compared to AVEC-2014 with MAE of 7.5 and 10.03 for both development and test respectively, and finally MAE proposed of 4.0 and 4.3 compared to AVEC-2016 of 5.36 and 5.72 for development and test parts respectively. Combining the newly developed features with state of the art spectral and voice quality features result in increasing the performance of the proposed models. The MAE (development, test) improved to (6.6, 6.8) for AVEC-2013, (7.1) for AVEC-2014 and (3.6, 4.0) for AVEC-2016. Based on these results, the new set of features demonstrated their ability in predicting the severity of depression when predicting the BDI and PHQ depression scores with better results compared to the baseline audio modalities and for all of the three datasets. In the same work, a depression language-agnostic system was constructed. This system combined samples from the training and development partitions and used them to train the proposed language-independent depression evaluation system, and only the test sets kept separated for the evaluation purpose. The results from using this design improved the MAE for AVEC-2013 and AVEC-2016 from 6.8 to 5.9 and 4.0 to 3.8 respectively; however, the result for the AVEC-2014 was 7.1 which is the same compared to the previous model. Even with the latter result, the improvement in AVEC-2013 and AVEC-2016 demonstrated the feasibility of using the newly developed speech activity features as an automatic and language agnostic depression assessment system. The results in this chapter provide the answer for the third and last research questions and are being prepared for a journal publication.

9.1.4 The feasibility of developing an objective tool for screening bipolar disorder?

Chapter 8 introduced an automatic screening system for bipolar states severity estimation. The model developed using the same set of features that were created in chapter 7. The proposed system tested with AVEC-2018 bipolar challenge dataset. In which three partitions were provided, training with 104, development with 60 and test set with 54 audio/video recordings. Both the training and development samples were labelled into three classes class1: remission, class2: hypo-mania and class3: mania using the young mania rating scales YMRS, however, the owner of the dataset kept the labels for the test set hidden from the public. This challenge will show how feasible is to develop bipolar screening tools to aid the clinicians in diagnosis. Several models were trained, and the best model was selected and achieved unweighted average recall UAR of 71.4% with the development set. This model was used to predict the labels for the test set, and the received result (from the dataset owners) showed and overfitting issue, as the test UAR was 42.59%. The problem with this dataset was the unbalance distribution of samples between the classes. To overcome this limitation and to produce a more generalised model that can perform similarly in both sets, a reduction on training set performed by moving samples from the class with the majority of samples to the development set. Although this method reduced the training size and the UAR on the development set to 66%, it improved the test set result to UAR of 53.7%, which outperformed both baseline audio and video modalities that have UAR of 50% and 46.3% respectively. This result again showing the usefulness of the newly proposed activity features in capturing not only depression behaviours but also discriminate the three severity levels of bipolar disease. This work answered the fourth and last research questions and will be reported in a journal paper.

.

9.2 Future work

9.2.1 Evaluate the dementia detection system with other datasets

Merhidari *et al.* [229] developed an avatar based system to screen patients at high risk of developing dementia. This experiment was conducted with group of patients participated in a study administered by the Royal Hallamshire hospital. The system proposed in chapter 6 will be evaluated with this dataset, and the result will be compared to Merhidari [229]. Future directions should also investigate larger numbers of participants with MCI who are in the earlier stages of AD [279] and aim to correlate noninvasively collected disease markers into an established tool for patients with cognitive concerns [280].

9.2.2 Improve the proposed MMSE monitoring system

The proposed system illustrated in chapter 5 that provides longitudinal MMSE scores monitoring needs to be improved. The system assumes a linear relationship between the AD progression and the changes in MMSE scores, while in the sever AD stage the expected MMSE scores will be rapidly decreased.

9.2.3 Extracting other type of features

There are extra acoustic features not investigated in this study due to time limits, for example, the Teager Energy Operator (TEO) which characterise the resonances in vocal tract produced by a nonlinear airflow in the cavity. TEO reported to be informative in detecting depression condition [12]; thus, it and other features might improve the performance.

9.2.4 Developing unified mental and mood disorders detection system

Based on the fact that there are overlapping symptoms between dementia and depression, the future direction will investigate the feasibility of developing unified dementia and depression screening tool, using the newly developed activity features and the models from this study. First is to create proper training, development and test partitions from all datasets used in this study including the DementiaBank, Royall Hallamshire and all AVEC challenge datasets. Secondly, is to develop a system capable of classifying between eight classes AD, MCI, HC, FMD, depression, remission, hypo-mania and mania. Also, to design a system to predict the scores of MMSE, BDI, PHQ, and YRMS for all dataset. Furthermore, to investigate how the clinicians could best utilise the proposed systems, what sort of audio-data acquisition devices are suitable for longitudinal monitoring.

9.2.5 Test the proposed models with datasets with different languages

In Iraq, the current tools and screening procedures for dementia and depression were either old or very limited and expensive; in fact, many patients travel to other countries to obtain a diagnosis for such conditions. Therefore, future work will be introducing the proposed systems to the doctors and planning for the possibility of utilising them in their screening procedure. The analysis will show how these systems performed using the Arabic language; also, the variety of conditions will allow the development of new models.

Bibliography

[1] A. Dowrick and A. Southern, *Dementia 2014: opportunity for change*. Alzheimer's Society, 2014.

[2] H. Goodglass, E. Kaplan, and S. Weintraub, *Boston naming test*. Lea & Febiger, 1983.

[3] A. L. Byers and K. Yaffe, "Depression and risk of developing dementia," *Nature Reviews Neurology*, vol. 7, no. 6, p. 323, 2011.

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

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

[6] A. D. Korczyn and I. Halperin, "Depression and dementia," *Journal of the neurological sciences*, vol. 283, no. 1-2, pp. 139–142, 2009.

[7] A. Alzheimer's, "2015 Alzheimer's disease facts and figures." *Alzheimer's & dementia: the journal of the Alzheimer's Association*, vol. 11, no. 3, p. 332, 2015.

[8] M. Kane and G. Terry, "Dementia 2015: Aiming higher to transform lives," *London: Alzheimer's Society*, 2015.

[9] M. Prince, R. Bryce, and C. Ferri, "World Alzheimer Report 2011: The benefits of early diagnosis and intervention," 2018.

[10] D. M. Holtzman, J. C. Morris, and A. M. Goate, “Alzheimer’s disease: the challenge of the second century,” *Science translational medicine*, vol. 3, no. 77, pp. 77sr1–77sr1, 2011.

[11] “Dementia: applying All Our Health,” https://www.gov.uk/government/publications/dementia-applying-all-our-health/dementia-applying-all-our-health, Online; accessed: 2019-05-30.

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

[13] “World Health Organization fact sheet,” https://www.who.int/news-room/fact-sheets/detail/depression, Online; accessed: 2019-07-18.

[14] World Health Organization, “Depression and other common mental disorders: global health estimates,” World Health Organization, Tech. Rep., 2017.

[15] P. McCrone, S. Dhanasiri, A. Patel, M. Knapp, and S. Lawton-Smith, “Paying the price,” 2008.

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

[17] J. L. Dieleman, R. Baral, M. Birger, A. L. Bui, A. Bulchis, A. Chapin, H. Hamavid, C. Horst, E. K. Johnson, J. Joseph *et al.*, “Us spending on personal health care and public health, 1996-2013,” *Jama*, vol. 316, no. 24, pp. 2627–2646, 2016.

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

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

[20] W. P. Suicide, “A global imperative,” *World Health Organization*, 2014.

[21] T. E. Joiner Jr, J. S. Brown, and L. R. Wingate, "The psychology and neurobiology of suicidal behavior," *Annu. Rev. Psychol.*, vol. 56, pp. 287–314, 2005.

[22] A. McGirr, J. Renaud, M. Seguin, M. Alda, C. Benkelfat, A. Lesage, and G. Turecki, "An examination of DSM-IV depressive symptoms and risk for suicide completion in major depressive disorder: a psychological autopsy study," *Journal of affective disorders*, vol. 97, no. 1-3, pp. 203–209, 2007.

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

[24] B. S. Diniz, M. A. Butters, S. M. Albert, M. A. Dew, and C. F. Reynolds, "Late-life depression and risk of vascular dementia and Alzheimer's disease: systematic review and meta-analysis of community-based cohort studies," *The British Journal of Psychiatry*, vol. 202, no. 5, pp. 329–335, 2013.

[25] J. Appell, A. Kertesz, and M. Fisman, "A study of language functioning in Alzheimer patients," *Brain and language*, vol. 17, no. 1, pp. 73–91, 1982.

[26] K. A. Bayles, "Language function in senile dementia," *Brain and language*, vol. 16, no. 2, pp. 265–280, 1982.

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

[28] S. A. Cooley, J. M. Heaps, J. D. Bolzenius, L. E. Salminen, L. M. Baker, S. E. Scott, and R. H. Paul, "Longitudinal change in performance on the montreal cognitive assessment in older adults," *The Clinical Neuropsychologist*, vol. 29, no. 6, pp. 824–835, 2015.

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

[30] J. Gatt, C. Nemeroff, C. Dobson-Stone, R. Paul, R. Bryant, P. Schofield, E. Gordon, A. Kemp, and L. Williams, "Interactions between BDNF Val66Met polymorphism and early life stress predict brain and arousal pathways to syndromal depression and anxiety," *Molecular psychiatry*, vol. 14, no. 7, p. 681, 2009.

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

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

[33] S. O. Orimaye, J. S.-M. Wong, and K. J. Golden, "Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 78–87.

[34] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[35] K. Sirts, O. Piguet, and M. Johnson, "Idea density for predicting Alzheimer's disease from transcribed speech," *arXiv preprint arXiv:1706.04473*, 2017.

[36] L. Hernández-Domínguez, S. Ratté, G. Sierra-Martínez, and A. Roche-Bergua, "Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 10, pp. 260–268, 2018.

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

[38] Alzheimer's Association, "2011 Alzheimer's disease facts and figures." *Alzheimer's & dementia: the journal of the Alzheimer's Association*, vol. 7, no. 2, p. 208, 2011.

[39] S. Ray, S. Davidson, and U. Age, “Dementia and cognitive decline,” *A reivew of the evidence*, 2014.

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

[41] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, “Ways toward an early diagnosis in Alzheimer’s disease: the Alzheimer’s disease neuroimaging initiative (ADNI),” *Alzheimer’s & Dementia*, vol. 1, no. 1, pp. 55–66, 2005.

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

[43] I. G. McKeith, “Dementia with Lewy bodies,” *The British Journal of Psychiatry*, vol. 180, no. 2, pp. 144–147, 2002.

[44] M. Fernández-Matarrubia, J. Matías-Guiu, T. Moreno-Ramos, and J. Matías-Guiu, “Behavioural variant frontotemporal dementia: Clinical and therapeutic approaches,” *Neurología (English Edition)*, vol. 29, no. 8, pp. 464–472, 2014.

[45] A. Alberdi, A. Aztiria, and A. Basarab, “On the early diagnosis of Alzheimer’s Disease from multimodal signals: A survey,” *Artificial Intelligence in Medicine*, vol. 71, pp. 1–29, 2016.

[46] D. E. Barnes, G. S. Alexopoulos, O. L. Lopez, J. D. Williamson, and K. Yaffe, “Depressive symptoms, vascular disease, and mild cognitive impairment: findings from the cardiovascular health study,” *Archives of general psychiatry*, vol. 63, no. 3, pp. 273–279, 2006.

[47] B. Klimova, P. Maresova, M. Valis, J. Hort, and K. Kuca, “Alzheimer’s disease and

language impairments: social intervention and medical treatment," *Clinical interventions in aging*, vol. 10, p. 1401, 2015.

[48] *Factsheet: The later stages of dementia.*, Alzheimer's Society, May 2017.

[49] B. McGleenon, K. Dynan, and A. Passmore, "Acetylcholinesterase inhibitors in Alzheimer's disease," *British journal of clinical pharmacology*, vol. 48, no. 4, p. 471, 1999.

[50] R. Cacabelos, M. Takeda, and B. Winblad, "The glutamatergic system and neurodegeneration in dementia: preventive strategies in alzheimer's disease," *International journal of geriatric psychiatry*, vol. 14, no. 1, pp. 3–47, 1999.

[51] NHS, "What are the treatments for dementia?" https://www.nhs.uk/conditions/dementia/treatment/, Online; accessed 27 January-2019.

[52] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia?" *Dementia and Geriatric Cognitive Disorders*, vol. 37, no. 5-6, pp. 327–334, 2014.

[53] S. Singh, R. S. Bucks, and J. M. Cuerden, "Evaluation of an objective technique for analysing temporal variables in dat spontaneous speech," *Aphasiology*, vol. 15, no. 6, pp. 571–583, 2001.

[54] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 2081–2090, 2011.

[55] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Toward the automation of diagnostic conversation analysis in patients with memory complaints," *Journal of Alzheimer's Disease*, vol. 58, no. 2, pp. 373–387, 2017.

[56] K. Schmidtke, S. Pohlmann, and B. Metternich, “The syndrome of functional memory disorder: definition, etiology, and natural course,” *The American Journal of Geriatric Psychiatry*, vol. 16, no. 12, pp. 981–988, 2008.

[57] B. A. Kramer, “Depressive pseudodementia,” *Comprehensive psychiatry*, vol. 23, no. 6, pp. 538–544, 1982.

[58] R. Camicioli, “Diagnosis and differential diagnosis of dementia,” *Dementia*, pp. 1–13, 2014.

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

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

[61] M. F. Folstein, S. E. Folstein, and P. R. McHugh, “Mini-mental state”: a practical method for grading the cognitive state of patients for the clinician,” *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, 1975.

[62] J. C. Anthony, L. LeResche, U. Niaz, M. R. Von Korff, and M. F. Folstein, “Limits of the ‘Mini-Mental State’as a screening test for dementia and delirium among hospital patients,” *Psychological medicine*, vol. 12, no. 2, pp. 397–408, 1982.

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

[64] E. Borland, K. Nägga, P. M. Nilsson, L. Minthon, E. D. Nilsson, and S. Palmqvist, “The montreal cognitive assessment: normative data from a large swedish population-based cohort,” *Journal of Alzheimer’s Disease*, vol. 59, no. 3, pp. 893–901, 2017.

[65] P. Mathuranath, P. Nestor, G. Berrios, W. Rakowicz, and J. Hodges, “A brief cognitive test battery to differentiate Alzheimer’s disease and frontotemporal dementia,” *Neurology*, vol. 55, no. 11, pp. 1613–1620, 2000.

[66] E. Mioshi, K. Dawson, J. Mitchell, R. Arnold, and J. R. Hodges, “The Addenbrooke’s Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening,” *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, vol. 21, no. 11, pp. 1078–1085, 2006.

[67] S. Hsieh, S. Schubert, C. Hoon, E. Mioshi, and J. R. Hodges, “Validation of the Addenbrooke’s Cognitive Examination III in frontotemporal dementia and Alzheimer’s disease,” *Dementia and geriatric cognitive disorders*, vol. 36, no. 3-4, pp. 242–250, 2013.

[68] O. Pedraza, B. C. Sachs, T. J. Ferman, B. K. Rush, and J. A. Lucas, “Difficulty and discrimination parameters of boston naming test items in a consecutive clinical series,” *Archives of Clinical Neuropsychology*, vol. 26, no. 5, pp. 434–444, 2011.

[69] D. Wechsler, “Wechsler adult intelligence scale–Fourth Edition (WAIS–IV),” *San Antonio, TX: NCS Pearson*, vol. 22, p. 498, 2008.

[70] ——, “Wechsler memory scale.” 1945.

[71] ——, *WMS-IV: Wechsler Memory Scale*. Pearson, 2009.

[72] P. de Zwart, B. Jeronimus, and P. De Jonge, “Empirical evidence for definitions of episode, remission, recovery, relapse and recurrence in depression: a systematic review,” *Epidemiology and psychiatric sciences*, pp. 1–19, 2018.

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

[74] M. JH Balsters, E. J Krahmer, M. GJ Swerts, and A. JJM Vingerhoets, “Verbal and nonverbal correlates for depression: a review,” *Current Psychiatry Reviews*, vol. 8, no. 3, pp. 227–234, 2012.

[75] "Prevention of mental disorders," https://www.who.int/mental_health/evidence/en/prevention_of_mental_disorders_sr.pdf, 2004, Online; accessed: 2018-2-10.

[76] E. J. Nestler, M. Barrot, R. J. DiLeone, A. J. Eisch, S. J. Gold, and L. M. Monteggia, "Neurobiology of depression," *Neuron*, vol. 34, no. 1, pp. 13–25, 2002.

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

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

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

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

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

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

[83] M. Blais and L. Baer, "Understanding rating scales and assessment instruments," in *Handbook of clinical rating scales and assessment in psychiatry and mental health*. Springer, 2009, pp. 1–6.

[84] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *Journal of neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

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

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

[87] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber *et al.*, “The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression,” *Biological psychiatry*, vol. 54, no. 5, pp. 573–583, 2003.

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

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

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

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

[92] R. Firestone, “Firestone assessment of self-destructive thoughts,” *San Antonio, TX: Psychological Corporation*, 1996.

[93] D. Maust, M. Cristancho, L. Gray, S. Rushing, C. Tjoa, and M. E. Thase, “Psychiatric rating scales,” in *Handbook of Clinical Neurology*. Elsevier, 2012, vol. 106, pp. 227–237.

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

[95] C. Cusin, H. Yang, A. Yeung, and M. Fava, “Chapter 2. rating scales for depression,” *Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health*, 2010.

[96] A. J. Rush, C. M. Gullion, M. R. Basco, R. B. Jarrett, and M. H. Trivedi, “The inventory of depressive symptomatology (IDS): psychometric properties,” *Psychological medicine*, vol. 26, no. 3, pp. 477–486, 1996.

[97] I. M. Cameron, J. R. Crawford, A. H. Cardy, S. W. du Toit, K. Lawton, S. Hay, K. Mitchell, S. Sharma, S. Shivaprasad, S. Winning *et al.*, “Psychometric properties of the quick inventory of depressive symptomatology (QIDS-SR) in uk primary care,” *Journal of psychiatric research*, vol. 47, no. 5, pp. 592–598, 2013.

[98] S. A. Montgomery and M. Åsberg, “A new depression scale designed to be sensitive to change,” *The British journal of psychiatry*, vol. 134, no. 4, pp. 382–389, 1979.

[99] J. B. Williams and K. A. Kobak, “Development and reliability of a structured interview guide for the Montgomery-åsberg Depression Rating Scale (SIGMA),” *The British Journal of Psychiatry*, vol. 192, no. 1, pp. 52–58, 2008.

[100] W. W. Zung, “A self-rating depression scale,” *Archives of general psychiatry*, vol. 12, no. 1, pp. 63–70, 1965.

[101] J. B. Gabrys and K. Peters, “Reliability, discriminant and predictive validity of the Zung Self-Rating Depression Scale,” *Psychological Reports*, vol. 57, no. 3_suppl, pp. 1091–1096, 1985.

[102] W. W. Zung, “A rating instrument for anxiety disorders.” *Psychosomatics: Journal of Consultation and Liaison Psychiatry*, 1971.

[103] A. T. Beck, N. Epstein, G. Brown, and R. A. Steer, “An inventory for measuring clinical anxiety: psychometric properties.” *Journal of consulting and clinical psychology*, vol. 56, no. 6, p. 893, 1988.

[104] K. Kroenke, R. L. Spitzer, and J. B. Williams, “The phq-9,” *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[105] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe, “A brief measure for assessing generalized anxiety disorder: the GAD-7,” *Archives of Internal Medicine*, vol. 166, no. 10, pp. 1092–1097, 2006.

[106] G. Hasler, W. C. Drevets, H. K. Manji, and D. S. Charney, “Discovering endophenotypes for major depression,” *Neuropsychopharmacology*, vol. 29, no. 10, p. 1765, 2004.

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

[108] J. J. Mann, “Neurobiology of suicidal behaviour,” *Nature Reviews Neuroscience*, vol. 4, no. 10, p. 819, 2003.

[109] M. K. Nock, G. Borges, E. J. Bromet, C. B. Cha, R. C. Kessler, and S. Lee, “Suicide and suicidal behavior,” *Epidemiologic reviews*, vol. 30, no. 1, pp. 133–154, 2008.

[110] Y. I. Sheline, “Neuroimaging studies of mood disorder effects on the brain,” *Biological psychiatry*, vol. 54, no. 3, pp. 338–352, 2003.

[111] K. C. Evans, D. D. Dougherty, M. H. Pollack, and S. L. Rauch, “Using neuroimaging to predict treatment response in mood and anxiety disorders,” *Annals of Clinical Psychiatry*, vol. 18, no. 1, pp. 33–42, 2006.

[112] P. E. Croarkin, A. J. Levinson, and Z. J. Daskalakis, “Evidence for GABAergic inhibitory deficits in major depressive disorder,” *Neuroscience & Biobehavioral Reviews*, vol. 35, no. 3, pp. 818–825, 2011.

[113] G. Hasler and G. Northoff, “Discovering imaging endophenotypes for major depression,” *Molecular psychiatry*, vol. 16, no. 6, p. 604, 2011.

[114] H. Ring and J. Serra-Mestres, “Neuropsychiatry of the basal ganglia,” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 72, no. 1, pp. 12–21, 2002.

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

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

[117] H. D. Schmidt, R. C. Shelton, and R. S. Duman, “Functional biomarkers of depression: diagnosis, treatment, and pathophysiology,” *Neuropsychopharmacology*, vol. 36, no. 12, p. 2375, 2011.

[118] E. Domenici, D. R. Willé, F. Tozzi, I. Prokopenko, S. Miller, A. McKeown, C. Brittain, D. Rujescu, I. Giegling, C. W. Turck *et al.*, “Plasma protein biomarkers for depression and schizophrenia by multi analyte profiling of case-control collections,” *PLoS one*, vol. 5, no. 2, p. e9166, 2010.

[119] M. Owens, J. Herbert, P. B. Jones, B. J. Sahakian, P. O. Wilkinson, V. J. Dunn, T. J. Croudace, and I. M. Goodyer, “Elevated morning cortisol is a stratified population-level biomarker for major depression in boys only with high depressive symptoms,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 9, pp. 3638–3643, 2014.

[120] R. M. Carney, K. E. Freedland, and R. C. Veith, “Depression, the autonomic nervous system, and coronary heart disease,” *Psychosomatic medicine*, vol. 67, pp. S29–S33, 2005.

[121] D. Schneider, C. Regenbogen, T. Kellermann, A. Finkelmeyer, N. Kohn, B. Derntl, F. Schneider, and U. Habel, “Empathic behavioral and physiological responses to dynamic stimuli in depression,” *Psychiatry research*, vol. 200, no. 2-3, pp. 294–305, 2012.

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

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

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

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

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

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

[128] E. Kraepelin, “Manic depressive insanity and paranoia,” *The Journal of Nervous and Mental Disease*, vol. 53, no. 4, p. 350, 1921.

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

[130] K. R. Scherer, “Vocal affect expression: A review and a model for future research.” *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[131] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, “Speaking in Alzheimer’s disease, is that an early sign? importance of changes in language abilities in Alzheimer’s disease,” *Frontiers in aging neuroscience*, vol. 7, p. 195, 2015.

[132] D. A. Snowdon, S. J. Kemper, J. A. Mortimer, L. H. Greiner, D. R. Wekstein, and W. R. Markesbery, "Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study," *Jama*, vol. 275, no. 7, pp. 528–532, 1996.

[133] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000.

[134] P. Garrard, L. M. Maloney, J. R. Hodges, and K. Patterson, "The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author," *Brain*, vol. 128, no. 2, pp. 250–260, 2004.

[135] F. Cuetos, J. C. Arango-Lasprilla, C. Uribe, C. Valencia, and F. Lopera, "Linguistic changes in verbal expression: a preclinical marker of Alzheimer's disease," *Journal of the International Neuropsychological Society*, vol. 13, no. 3, pp. 433–439, 2007.

[136] L. J. Clark, M. Gatz, L. Zheng, Y.-L. Chen, C. McCleary, and W. J. Mack, "Longitudinal verbal fluency in normal aging, preclinical, and prevalent Alzheimer's disease," *American Journal of Alzheimer's Disease & Other Dementias®*, vol. 24, no. 6, pp. 461–468, 2009.

[137] X. Le, I. Lancashire, G. Hirst, and R. Jokel, "Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists," *Literary and linguistic computing*, vol. 26, no. 4, pp. 435–461, 2011.

[138] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp, "Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech," in *IEEE International Conference Mechatronics and Automation, 2005*, vol. 3. IEEE, 2005, pp. 1569–1574.

[139] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of sponta-

neous speech," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 27–37.

[140] P. Garrard, V. Rentoumi, B. Gesierich, B. Miller, and M. L. Gorno-Tempini, "Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse," *Cortex*, vol. 55, pp. 122–129, 2014.

[141] V. Rentoumi, L. Raoufian, S. Ahmed, C. A. de Jager, and P. Garrard, "Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology," *Journal of Alzheimer's Disease*, vol. 42, no. s3, pp. S3–S17, 2014.

[142] E. Prud'hommeaux and B. Roark, "Graph-based word alignment for clinical language evaluation," *Computational Linguistics*, vol. 41, no. 4, pp. 549–578, 2015.

[143] M. Asgari, J. Kaye, and H. Dodge, "Predicting mild cognitive impairment from spontaneous spoken utterances," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 3, no. 2, pp. 219–228, 2017.

[144] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert *et al.*, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112–124, 2015.

[145] K. Rockwood, J. Graham, and S. Fay, "Goal setting and attainment in Alzheimer's disease patients treated with donepezil," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 73, no. 5, pp. 500–507, 2002.

[146] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease," *Brain*, vol. 136, no. 12, pp. 3727–3737, 2013.

[147] "Dementia talkbank," https://dementia.talkbank.org/, online; accessed: 2015-09-07.

[148] S. O. Orimaye, J. S. Wong, K. J. Golden, C. P. Wong, and I. N. Soyiri, "Predicting probable Alzheimer's disease using linguistic deficits and biomarkers," *BMC bioinformatics*, vol. 18, no. 1, p. 34, 2017.

[149] L. Zhou, K. C. Fraser, and F. Rudzicz, "Speech recognition in Alzheimer's disease and in its assessment." in *Interspeech*, 2016, pp. 1948–1952.

[150] R. B. Ammar and Y. B. Ayed, "Speech processing for early Alzheimer Disease diagnosis: Machine learning based approach," in *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*. IEEE, 2018, pp. 1–8.

[151] B. MacWhinney, "The talkbank project," in *Creating and digitizing language corpora*. Springer, 2007, pp. 163–180.

[152] P. Klumpp, J. Fritsch, and E. Noeth, "ANN-based Alzheimer's disease classification from bag of words," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–4.

[153] J. Haiman, *Iconicity in syntax: proceedings of a Symposium on iconicity in syntax, Stanford, June 24-6, 1983*. John Benjamins Publishing, 1985, vol. 6.

[154] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Detecting signs of dementia using word vector representations." in *Interspeech*, 2018, pp. 1893–1897.

[155] J. Fritsch, S. Wankerl, and E. Nöth, "Automatic diagnosis of Alzheimer's disease using neural network language models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5841–5845.

[156] K. López-de Ipiña, J. B. Alonso, N. Barroso, M. Faundez-Zanuy, M. Ecay, J. Solé-Casals, C. M. Travieso, A. Estanga, and A. Ezeiza, "New approaches for Alzheimer's disease diagnosis based on automatic spontaneous speech analysis and emotional temperature," in *International Workshop on Ambient Assisted Living*. Springer, 2012, pp. 407–414.

[157] K. López-de Ipiña, J.-B. Alonso, C. M. Travieso, J. Solé-Casals, H. Egiraun, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage *et al.*, "On the selection

of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis," *Sensors*, vol. 13, no. 5, pp. 6730–6745, 2013.

[158] K. López-de Ipina, J. Solé-Casals, H. Eguiraun, J. B. Alonso, C. M. Travieso, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage, and B. Beitia, "Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach," *Computer Speech & Language*, vol. 30, no. 1, pp. 43–60, 2015.

[159] P. Boersma, "Praat: a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[160] L. Tóth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatlóczki, Z. Bánréti, M. Pákáski, and J. Kálmán, "A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech," *Current Alzheimer Research*, vol. 15, no. 2, pp. 130–138, 2018.

[161] A. Satt, A. Sorin, O. Toledo-Ronen, O. Barkan, I. Kompatsiaris, A. Kokonozi, and M. Tsolaki, "Evaluation of speech-based protocol for detection of early-stage dementia." in *INTERSPEECH*, 2013, pp. 1692–1696.

[162] J. Weiner, C. Herff, and T. Schultz, "Speech-Based Detection of Alzheimer's Disease in Conversational German." in *INTERSPEECH*, 2016, pp. 1938–1942.

[163] J. Weiner, M. Engelbart, and T. Schultz, "Manual and automatic transcriptions in dementia detection from speech." in *INTERSPEECH*, 2017, pp. 3117–3121.

[164] J. Weiner, M. Angrick, S. Umesh, and T. Schultz, "Investigating the effect of audio duration on dementia detection using acoustic features." in *Interspeech*, 2018, pp. 2324–2328.

[165] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[166] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech

using ASR and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.

[167] K. C. Fraser, K. L. Fors, and D. Kokkinakis, "Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment," *Computer Speech & Language*, vol. 53, pp. 121–139, 2019.

[168] A. Wallin, A. Nordlund, M. Jonsson, K. Lind, Å. Edman, M. Göthlin, J. Stålhammar, M. Eckerström, S. Kern, A. Börjesson-Hanson *et al.*, "The gothenburg mci study: design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up," *Journal of Cerebral Blood Flow & Metabolism*, vol. 36, no. 1, pp. 114–131, 2016.

[169] K. Cromnow and T. Landberg, *Skriftliga beskrivningar av bilden "Kakstölden". Insamling av referensvärden från friska försökspersoner.* Inst lingvisitk, Stockholms universitet, 2009.

[170] M. Yancheva, K. Fraser, and F. Rudzicz, "Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 134–139.

[171] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[172] A. C. Trevino, T. F. Quatieri, and N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, p. 42, 2011.

[173] E. Moore II, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE transactions on biomedical engineering*, vol. 55, no. 1, pp. 96–107, 2007.

[174] L.-S. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, "Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 5154–5157.

[175] K. E. B. Ooi, M. Lech, and N. B. Allen, "Multichannel weighted speech classification system for prediction of major depression in adolescents," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 2, pp. 497–506, 2012.

[176] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 142–150, 2012.

[177] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting depression: a comparison between spontaneous and read speech," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7547–7551.

[178] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[179] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 41–48.

[180] H. Meng, D. Huang, H. Wang, H. Yang, M. Ai-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 21–30.

[181] I. Bloch, "Information combination operators for data fusion: a comparative review with classification," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 26, no. 1, pp. 52–67, 1996.

[182] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, “Diagnosis of depression by behavioural signals: a multimodal approach,” in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 11–20.

[183] N. Cummins, V. Sethu, J. Epps, and J. Krajewski, “Probabilistic acoustic volume analysis for speech affected by depression,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

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

[185] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[186] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker, “Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression,” *depression*, vol. 1, no. 1, 2014.

[187] O. Simantiraki, P. Charonyktakis, A. Pampouchidou, M. Tsiknakis, and M. Cooke, “Glottal source features for automatic speech-based depression assessment.” in *INTERSPEECH*, 2017, pp. 2700–2704.

[188] V. Mitra and E. Shriberg, “Effects of feature type, learning algorithm and speaking style for depression detection from speech,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4774–4778.

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

[190] M. R. Morales and R. Levitan, “Speech vs. text: A comparative analysis of features for depression detection systems,” in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 136–143.

[191] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in opensmile, the munich open-source multimedia feature extractor,” in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM ’13. New York, NY, USA: ACM, 2013, pp. 835–838. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502224

[192] J. W. Pennebaker, M. E. Francis, and R. J. Booth, “Linguistic inquiry and word count: LIWC 2001,” *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[193] A. N. Rafferty and C. D. Manning, “Parsing three German treebanks: Lexicalized and unlexicalized baselines,” in *Proceedings of the Workshop on Parsing German*. Association for Computational Linguistics, 2008, pp. 40–46.

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

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

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

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

[198] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, “Multimodal measurement of depression using deep learning models,” in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 53–59.

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

[200] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, “Detecting depression with audio/text sequence modeling of interviews.” in *Interspeech*, 2018, pp. 1716–1720.

[201] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, “Depaudionet: An efficient deep model for audio based depression classification,” in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 35–42.

[202] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, “The natural history of Alzheimer’s disease: description of study cohort and accuracy of diagnosis,” *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[203] J. R. Booth, B. Mac Whinney, and Y. Harasaki, “Developmental differences in visual and auditory processing of complex sentences,” *Child Development*, vol. 71, no. 4, pp. 981–1003, 2000.

[204] D. Mazzoni, “Audacity®,” https://audacityteam.org/, 1999-2018.

[205] K. Lopez-de Ipina, J. B. Alonso, C. M. Travieso, H. Egiraun, M. Ecay, A. Ezeiza, N. Barroso, and P. Martinez-Lage, “Automatic analysis of emotional response based on nonlinear speech modeling oriented to Alzheimer disease diagnosis,” in *Intelligent Engineering Systems (INES), 2013 IEEE 17th International Conference on*. IEEE, 2013, pp. 61–64.

[206] A. De Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[207] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, “Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson’s disease symptom severity,” *Journal of the Royal Society Interface*, vol. 8, no. 59, pp. 842–855, 2011.

[208] K. Lopez-de Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martinez-Lage, and H. Eguiraun, “On automatic diagnosis of Alzheimer’s disease based on spontaneous speech analysis and emotional temperature,” *Cognitive Computation*, vol. 7, no. 1, pp. 44–55, 2015.

[209] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, “Accurate telemonitoring of Parkinson’s disease progression by noninvasive speech tests,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010.

[210] F. Gayraud, H.-R. Lee, and M. Barkat-Defradas, “Syntactic and lexical context of pauses and hesitations in the discourse of Alzheimer patients and healthy elderly subjects,” *Clinical linguistics & phonetics*, vol. 25, no. 3, pp. 198–209, 2011.

[211] T. Khan, J. Westin, and M. Dougherty, “Classification of speech intelligibility in parkinson’s disease,” *Biocybernetics and Biomedical Engineering*, vol. 34, no. 1, pp. 35–45, 2014.

[212] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” in *Readings in speech recognition*. Elsevier, 1990, pp. 65–74.

[213] P. N. Le, E. Ambikairajah, J. Epps, V. Sethu, and E. H. Choi, “Investigation of spectral centroid features for cognitive load classification,” *Speech Communication*, vol. 53, no. 4, pp. 540–551, 2011.

[214] K. K. Paliwal, “Spectral subband centroid features for speech recognition,” in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, pp. 617–620.

[215] B. Gajic and K. K. Paliwal, “Robust feature extraction using subband spectral centroid histograms,” in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP’01). 2001 IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 85–88.

[216] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

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

[218] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[219] I. Corp, “Ibm spss statistics for windows, version 25.0,” *Armonk, NY: IBM Corp*, 2017.

[220] S. S. Shapiro and M. B. Wilk, “An analysis of variance test for normality (complete samples),” *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

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

[222] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, “Novel speech signal processing algorithms for high-accuracy classification of Parkinson’s disease,” *IEEE transactions on biomedical engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[223] S. Luz, “Longitudinal monitoring and detection of alzheimer’s type dementia from spontaneous speech data,” in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2017, pp. 45–46.

[224] D. Hakkani-Tür, D. Vergyri, and G. Tur, "Speech-based automated cognitive status assessment," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[225] J. J. Meilán, F. Martínez-Sánchez, J. Carro, J. A. Sánchez, and E. Pérez, "Acoustic markers associated with impairment in language processing in Alzheimer's disease," *The Spanish journal of psychology*, vol. 15, no. 2, pp. 487–494, 2012.

[226] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

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

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

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

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

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

[232] J. C. Raven, *Guide to using the coloured progressive matrices.* HK Lewis & Co., 1958.

[233] J. R. Stroop, "Studies of interference in serial verbal reactions." *Journal of Experimental Psychology*, vol. 18, no. 6, p. 643, 1935.

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

[235] D. Wechsler, “Wechsler adult intelligence scale–Fourth Edition (WAIS–IV),” *San Antonio, Texas: Psychological Corporation*, 2014.

[236] R. C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, and L. Fratiglioni, “Mild cognitive impairment: a concept in evolution,” *Journal of Internal Medicine*, vol. 275, no. 3, pp. 214–228, 2014.

[237] K. Rascovsky, J. R. Hodges, D. Knopman, M. F. Mendez, J. H. Kramer, J. Neuhaus, J. C. Van Swieten, H. Seelaar, E. G. Dopper, C. U. Onyike *et al.*, “Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia,” *Brain*, vol. 134, no. 9, pp. 2456–2477, 2011.

[238] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen *et al.*, “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: the journal of the Alzheimer’s Association*, vol. 7, no. 3, pp. 270–279, 2011.

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

[240] S. Al-Hameed, M. Benaissa, and H. Christensen, “Simple and robust audio-based detection of biomarkers for Alzheimer’s disease,” in *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2016, pp. 32–36.

[241] ——, “Detecting and predicting alzheimer’s disease severity in longitudinal acoustic data,” in *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*. ACM, 2017, pp. 57–61.

[242] K. Bennys, G. Rondouin, C. Vergnes, and J. Touchon, “Diagnostic value of quantitative EEG in Alzheimer’s disease,” *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 31, no. 3, pp. 153 – 160, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0987705301002544

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

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

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

[246] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, “AVEC 2014: 3d dimensional affect and depression recognition challenge,” in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC ’14. New York, NY, USA: ACM, 2014, pp. 3–10. [Online]. Available: http://doi.acm.org/10.1145/2661806.2661807

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

[248] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. Narayanan, “Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist,” in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[249] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani,

F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[250] H.-y. Lee, T.-y. Hu, H. Jing, Y.-F. Chang, Y. Tsao, Y.-C. Kao, and T.-L. Pao, "Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition." in *INTERSPEECH*, 2013, pp. 215–219.

[251] D. Bone, M. P. Black, A. Ramakrishna, R. Grossman, and S. S. Narayanan, "Acoustic-prosodic correlates ofawkward'prosody in story retellings from adolescents with autism," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[252] S. Al-Hameed, M. Benaissa, H. Christensen, B. Mirheidari, D. Blackburn, and M. Reuber, "A new diagnostic approach for the identification of patients with neurodegenerative cognitive complaints," *PLOS ONE*, vol. 14, no. 5, pp. 1–18, 05 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0217388

[253] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, p. e0144610, 2015.

[254] T. Giannakopoulos and A. Pikrakis, *Introduction to audio analysis: a MATLAB® approach*. Academic Press, 2014.

[255] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006.

[256] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[257] E. Severus and M. Bauer, "Diagnosing bipolar disorders in DSM-5," 2013.

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

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

[260] A. Guidi, J. Schoentgen, G. Bertschy, C. Gentili, E. P. Scilingo, and N. Vanello, "Features of vocal frequency contour and speech rhythm in bipolar disorder," *Biomedical Signal Processing and Control*, vol. 37, pp. 23–31, 2017.

[261] A. Guidi, E. P. Scilingo, C. Gentili, G. Bertschy, L. Landini, and N. Vanello, "Analysis of running speech for the characterization of mood state in bipolar patients," in *2015 AEIT International Annual Conference (AEIT)*. IEEE, 2015, pp. 1–6.

[262] A. Maxhuni, A. Muñoz-Meléndez, V. Osmani, H. Perez, O. Mayora, and E. F. Morales, "Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients," *Pervasive and Mobile Computing*, vol. 31, pp. 50–66, 2016.

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

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

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

[266] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emo-

tional speech," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 478–484.

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

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

[269] kristopher kyle, "Suite of Automatic Linguistic Analysis Tools (SALAT)," http://www.kristopherkyle.com/, [Online].

[270] M. T. C. Ltd, "Face++," https://www.faceplusplus.com/, [Online].

[271] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

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

[273] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

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

[275] D. Ververidis and C. Kotropoulos, "Fast sequential floating forward selection applied to emotional speech features estimated on des and susas data collections," in *2006 14th European Signal Processing Conference*. IEEE, 2006, pp. 1–5.

[276] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, “COVAREP—A collaborative voice analysis repository for speech technologies,” in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

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

[278] W. Zong, G.-B. Huang, and Y. Chen, “Weighted extreme learning machine for imbalance learning,” *Neurocomputing*, vol. 101, pp. 229–242, 2013.

[279] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, “Mild cognitive impairment: clinical characterization and outcome,” *Archives of Neurology*, vol. 56, no. 3, pp. 303–308, 1999.

[280] B. Dubois, H. H. Feldman, C. Jacova, H. Hampel, J. L. Molinuevo, K. Blennow, S. T. DeKosky, S. Gauthier, D. Selkoe, R. Bateman *et al.*, “Advancing research diagnostic criteria for Alzheimer’s disease: the IWG-2 criteria,” *The Lancet Neurology*, vol. 13, no. 6, pp. 614–629, 2014.