Machine Learning on Speech Audio for Extracting Indicators of Psychiatric and Neurodegenerative Conditions

by

Mashrura Tasnim

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science
University of Alberta

Abstract

The objective of this thesis is to develop and validate computational models of vocal expressions to predict the severity of (i) psychiatric disorders, such as depression, anxiety, and stress, and (ii) neurodegenerative diseases, like Alzheimer's Dementia (AD) and mild cognitive impairment (MCI). The fundamental assumption of this work is that the above conditions, and possibly others, impact the individual's ability to produce language, and therefore their vocal expressions are distinguishable from those of healthy individuals.

In the quest to better understand and predict various aspects of these conditions, this thesis explores a comprehensive exploration of vocal expressions. Leveraging a variety of datasets spanning psychiatric disorders like depression and anxiety and neurodegenerative diseases such as AD and MCI, this research aims to decode the intricate nuances embedded within vocal tones.

The methodological approach incorporates sophisticated audio analysis techniques with supervised machine learning models trained on labelled speech samples. Through a meticulously designed pipeline, encompassing noise reduction, feature extraction, and model training, the research endeavours to establish connections between vocal cues and mental and cognitive health conditions. This thesis underscores the potential of audiovisual cues as invaluable markers for advancing our comprehension and prediction of mental health conditions and cognitive competencies, thereby paving the way for more effective diagnostic and intervention strategies.

Preface

This doctoral thesis represents the culmination of years of dedicated research, experimentation, and collaboration in the area of application of artificial intelligence and machine learning in mental and cognitive health. The journey narrated in this thesis reflects not only my personal academic growth but also the invaluable guidance and support of my mentors, colleagues, and collaborators.

Chapter 2 of this thesis builds upon research initially presented at the 32nd Canadian Conference on Artificial Intelligence (Canadian AI 2019). As the primary author of this work, conducted under the supervision of Prof. Eleni Stroulia, I undertook the design and execution of the experiments detailed therein.

Chapter 3 contains unpublished experimental methodology which was submitted to IEEE Access, conducted with Prof. Stroulia's guidance, further expanding upon the foundation laid in earlier chapters.

Chapter 4 draws upon research conducted during an internship at Winterlight Labs (acquired by Cambridge Cognition in 2023), under the supervision of Jekaterina Novikova. The findings presented were initially disseminated at the 21st IEEE International Conference on Machine Learning and Applications (ICMLA) in 2022.

The newly curated dataset presented in Chapter 5 and the preliminary analysis on it have been published in the 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). As the first author, I conducted 70% tasks in the experiments and documentation with the assistance of Ramon Diaz Ramos, under the supervision of Prof. Eleni Stroulia; in collaboration with Prof. Luis A. Trejo at the Tecnologico de Monterrey, Atizapán, Mexico. The study protocol was approved by Alberta Research Information Services (ARISE), project name "Physiological and Behavioural Data Collection to Predict Anxiety, Stress, and Depression Among Students", No. Pro00116909, January 24, 2022.

The analysis done in Chapter 6 was published in the Frontiers in Computer Science (2021), describing our methodology for competing in the ADReSS 2020 challenge. We participated in this challenge as a team consisting of myself, Zehra Shah, Jeffrey Sawalha, and Shi-ang Qi. Prof. Eleni Stroulia and Prof. Russell Greiner jointly supervised the team. As a participating team member, I worked on the acoustic pipeline, which is approximately 30% of the experimentation. I also compiled about 30% of the documentation.

The experimental methodology to address the TAUKADIAL 2024 challenge presented in Chapter 7 has been submitted to INTERSPEECH 2024, Kos Island, Greece, and is currently under review. The University of Alberta team participating in this challenge included myself, Shi-ang Qi, Fei Wang, Mahtab Farrokh, Siyang Tian, Jimuel Jr. Celeste under the guidance of Prof. Eleni Stroulia and Prof. Russell Greiner. As a member of the participating team and the first author of the paper, I conducted approximately 40% of the experimentation and 70% of the documentation.

Throughout this thesis, my aim is not only to present findings but also to contribute to the collective body of knowledge in computational psychiatry. I am profoundly grateful to all who have contributed to this journey, and I hope that this work serves as a testament to the power of collaboration, perseverance, and intellectual curiosity in advancing scientific understanding.

Acknowledgements

I want to express my deepest gratitude to my esteemed supervisor Prof. Eleni Stroulia, for her unwavering support, invaluable guidance, and scholarly insight throughout this doctoral journey. Her mentorship has been instrumental in shaping not only the research presented in this thesis but also my academic and personal growth.

I am profoundly thankful to my collaborators and colleagues, whose contributions have enriched this work immeasurably. Special thanks to Prof. Russell Greiner, Prof. Luis A. Trejo, Jekaterina Novikova, Ramon Diaz Ramos, Zehra Shah, Jeffrey Sawalha, Shi-ang Qi, Fei Wang, Mahtab Farrokh, Siyang Tian, and Jimuel Jr Celeste for their dedication, expertise, and collaborative spirit.

I am also indebted to my family for their unwavering support, encouragement, and understanding throughout this journey. Their love, patience, and belief in me have been a constant source of inspiration and motivation.

Lastly, I extend my heartfelt thanks to all the researchers, educators, and institutions whose work has paved the way for mine. It is through their collective efforts that the boundaries of knowledge continue to expand, and I am honoured to contribute to this ongoing pursuit.

Table of Contents

List of Tables

List of Figures

Abbreviations

AD: Alzheimer's disease.

AVEC: Audio Visual Emotion-recognition Challenge.

CNN: Convolutional neural network.

DEPAC: DEPression and Anxiety Crowdsourced corpus.

DT: Decision tree.

FNN: Feedforward neural network.

GBT: Gradient boosting tree.

HMM: Hidden Markov model.

LR: Logistic regression.

NN: Neural network.

PCA: Principal component analysis.

RF: Random forest.

SVM: Support vector machine.

XGB: Extreme gradient booting.

Chapter 1

Introduction

The global burden of mental health disorders and cognitive decline is substantial, with statistics indicating a pressing need for innovative solutions. According to the World Health Organization (WHO), depression affects over 280 million people worldwide [1], while Alzheimer's disease and other forms of dementia impact an estimated 55 million individuals [2]. Moreover, the prevalence of these conditions is expected to rise significantly in the coming years, posing challenges to healthcare systems worldwide [3]. Compounding this issue is the shortage of resources and trained professionals, particularly in low- and middle-income countries. The scarcity of specialists often results in delayed diagnosis and inadequate treatment for individuals suffering from these disorders. Consequently, there is a growing demand for digital solutions that could enable more timely and accessible support to those in need.

In response to these challenges, researchers have increasingly turned to speech and vocal tone-oriented systems as promising avenues for digital monitoring and assessment of mental health conditions and cognitive impairment, especially early, when treatment is more effective. Speech analysis offers a wealth of information about the severity and progression of various illnesses, including Alzheimer's disease, frontotemporal dementia, Parkinson's disease, schizophrenia, and depression. For instance, individuals with Alzheimer's disease often exhibit errors in speech, such as using incorrect or meaningless words, along with increased pauses [4]. People with schizophrenia tend to have disorganized and sometimes unintelligible speech [5]. People with Parkinson's disease may speak softly and display limited emotional expression [6]. Depression can manifest in the vocal tract and articulatory changes, resulting in monotone and less variable speech [7].

Leveraging automated speech analysis for the detection and assessment of these conditions offers several advantages over traditional methods. Firstly, it aligns with the ecologically valid nature of speech, as communication is an inherent aspect of human interaction and can be seamlessly integrated into daily routines without imposing significant burdens or complexity. Secondly, speech-based assessment systems hold functional relevance, as effective communication is essential for various aspects of daily life. Building upon existing research and studies in the field of speech-based mental health assessment, this thesis endeavors to develop a non-invasive monitoring and support system that capitalizes on the rich information embedded within speech signals. Through a phased approach, this thesis aims to advance digital solutions for mental health care (Figure 1.1). The proposed system will capture audio recordings through apps like digital diaries and transmit them to a secure server after encryption. A trained machine learning model will then analyze the recordings to infer possible indications of mental disorders and their severity. From a wide array of possible interventions, the system will suggest the one tailored to the individual's needs, ultimately enhancing accessibility and efficacy in diagnosis and intervention.

In the quest to better understand and predict various aspects of mental health conditions and cognitive competency, this thesis explores a comprehensive exploration of vocal expressions. Leveraging a variety of datasets spanning psychiatric disorders like depression and anxiety and neurodegenerative diseases such as Alzheimer's dementia (AD) and mild cognitive impairment (MCI), this research aims to decode the intricate nuances embedded within vocal tones. The methodological approach incorporates sophisticated audio analysis techniques with supervised machine learning models trained on labeled speech samples. Through a meticulously designed pipeline,

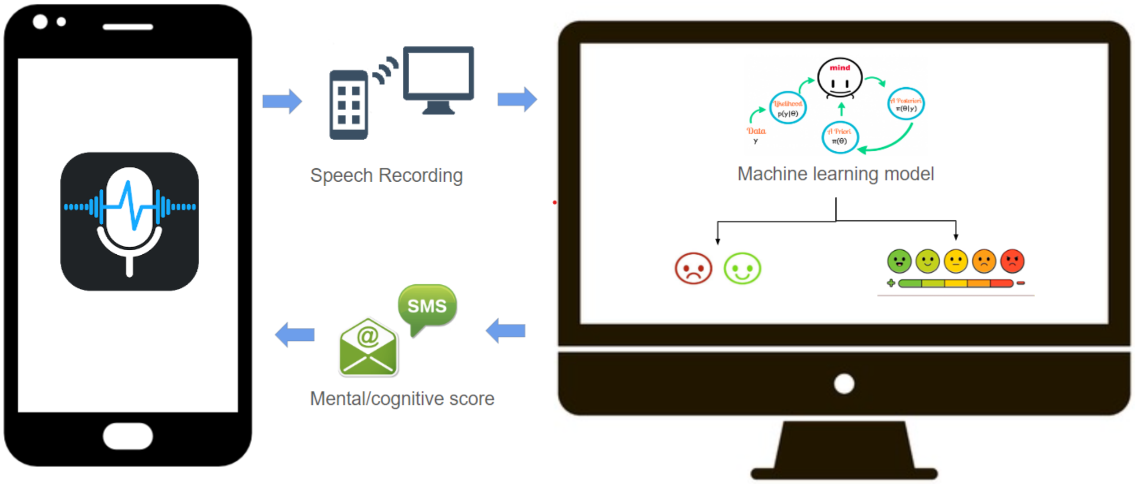


Figure 1.1: Envisioned speech-based monitoring and support system

encompassing noise reduction, feature extraction, and model training, the research endeavors to establish connections between vocal cues and mental and cognitive health conditions. This thesis underscores the potential of audiovisual cues as invaluable markers for advancing our comprehension and prediction of mental health conditions and cognitive competencies, thereby paving the way for more effective diagnostic and intervention strategies.

1.1 Objective

The objective of this thesis is to develop and validate computational models of vocal expressions to predict the severity of (i) psychiatric disorders, such as depression, anxiety, and stress, and (ii) neurodegenerative diseases, like AD and MCI. The fundamental assumption of this work is that the above conditions, and possibly others, impact the individual's ability to produce language, and therefore their vocal expressions are distinguishable from those of healthy individuals.

1.2 Data

This thesis systematically analyzes a wide variety of datasets. The datasets cover multiple modalities, including audio and text, and are annotated with subjective and objective measures of health conditions, assessed by healthcare experts and self-reported questionnaires respectively. The content of the datasets ranges from structured and semi-structured interviews, guided reading, and spontaneous speech. Table 1.1 provides an overview of the datasets used in the experiments reported in this thesis.

Table 1.1: The Datasets

	AVEC 2013	**AVEC 2017**	**AVEC 2019**	**DEPAC**	**YouthDASS (Depression)**	**YouthDASS (Anxiety)**	**YouthDASS (Stress)**
Language	German	English	English	English	English and Spanish	English and Spanish	English and Spanish
Format	Guided reading and freeform speech	Semi-structured interview	Semi-structured interview	Combination of guided reading and freeform speech tasks	Combination of guided reading and freeform speech tasks	Combination of guided reading and freeform speech tasks	Combination of guided reading and freeform speech tasks
Length of samples	< 5 minutes	5 to 20 minutes	15 to 25 minutes	5 to 50 seconds	23 to 67 seconds	23 to 67 seconds	23 to 67 seconds
#Recordings per session	2	1	1	5	2	2	2
#Samples	300	189	275	2,765	1024	1024	1024
#Training/development /test samples	100/100/100	107/35/47	163/56/56	Nested 10-fold CV	Nested 5-fold CV	Nested 5-fold CV	Nested 5-fold CV
Scale (Range)	BDI-II (0-63)	PHQ-8(0-24)	PHQ-8(0-24)	PHQ-9(0-27)	DASS-21 (0-42)	DASS-21 (0-42)	DASS-21 (0-42)
Threshold	19	10	10	10	9	7	15
%Samples above threshold (Patients)	36	23.94	24.2	15.12	22.59	23.74	15.35
Average score (SD)	15.3 (SD = 12.3)	6.67 (SD = 5.75)	6.64 (SD = 5.99)	6.56 (SD = 5.56)	5.94 (SD = 7.55)	4.79 (SD = 6.90)	7.12 (SD = 8.20)

- **Benchmark depression datasets:** The depression studies are initiated with benchmark datasets such as the Audio Visual Emotion-recognition Challenge (AVEC) 2013, AVEC 2017, and AVEC 2019. AVEC 2013 depression corpus consists of 300 German speech samples conducting guided speech tasks labeled with BDI-II scores (range 0 to 63). AVEC 2017 and AVEC 2019 depression datasets comprise semi-structured clinical interviews in English paired with PHQ-8 (range 0-24) scores.

- **DEPAC corpus:** DEPression and Anxiety Crowdsourced corpus (DEPAC) offers a glimpse into the complexities of mental health within naturalistic contexts. It consists of 2,674 audio samples collected from 571 English-speaking subjects located in Canada and the United States, aged between 18 and 76 years. The data was collected via crowdsourcing and consists of a variety of self-administered speech tasks (Table 1.2). The speech tasks were curated to increase the phonemic variety and were supported by literature on detecting mental disorders, such as AD [8] and depression [9], [10], [11] from speech. The participants completed these tasks using Amazon Mechanical Turk (mTurk) in their native environment. This dataset includes depression and anxiety scores measures on PHQ-9 and GAD-7 scales, respectively.

Table 1.2: Speech tasks in the DEPAC corpus

Speech Task	**Description**	**Average Duration**
Phoneme Task	Record “aah” sound for as long as the participant could hold breath	5.79 sec
Phonemic Fluency	Pronounce as many unique words as possible starting with the letters “F”, “A” or “S”	22.13 sec
Picture Description	Describe a picture shown on the screen	46.60 sec
Semantic Fluency	Describe a positive experience they expected to have within five years in future	43.76 sec
Prompted Narrative	Tell a personal story, describing the day, a hobby, or a travel experience	45.34 sec

- **YouthDASS dataset:** The analysis extends to the longitudinal multilingual YouthDASS dataset, shedding light on the temporal dynamics of depression, anxiety, and stress. To curate this dataset, English and Spanish-speaking participants are recruited among interested undergraduate and graduate students of the University of Alberta, Canada, and Tec de Monterrey, Mexico to record speech samples twice a week for three months. In each recording session, the participants perform a free-form speech task describing a memorable event, a hobby, or a favourite person, along with a guided reading task. At the end of each recording session, the participants fill out a DASS-21 questionnaire as a measure of their depression, anxiety, and stress; three commonly comorbid disorders. 1,049 samples (838 Spanish, and 211 English) are collected from the participants' mobile devices using an Android application. The study protocol (Pro00116909) is approved by Alberta Research Information Services (ARISE).

- **ADReSS 2020:** In Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge dataset, the subjects describe the 'cookie theft' picture in English. The dataset contains speech samples from 108 subjects (54 AD patients, 54 healthy) with corresponding manually annotated transcripts and MMSE scores. 48 samples are withheld by the organizers to evaluate the predictions.

- **TAUKADIAL 2024:** In the recent Speech-Based Cognitive Assessment in Chinese and English challenge (TAUKADIAL 2024), the challenge corpus includes three samples for each subject performing picture description tasks in either English or Chinese. 387 samples are provided for training (165 healthy, 222 MCI), while 120 are held out for evaluation.

  In both ADReSS 2020 and TAUKADIAL 2024 datasets, the cognitive health status of the subjects is certified by healthcare professionals, while the disorder severity is measured in MMSE score (Range 0-30, healthy if > 24). The mean MMSE scores in the two datasets are 23.08 (SD = 7.13) and 27.22 (SD = 3.32) respectively.

Through comprehensive examinations of these datasets, the thesis aims to construct a robust foundation for the development of non-invasive speech-based mental and cognitive health monitoring systems, capable of addressing the multifaceted challenges posed by these debilitating conditions.

1.3 Audio Analysis for Predicting Mental Health Conditions

The thesis describes the development of a speech analysis pipeline to analyze the speech corpus to infer mental disorders and cognitive impairment. The pipeline has three major components, as illustrated in Figure 1.2: (i) data preprocessing, (ii) machine-learning model construction, and (iii) prediction.

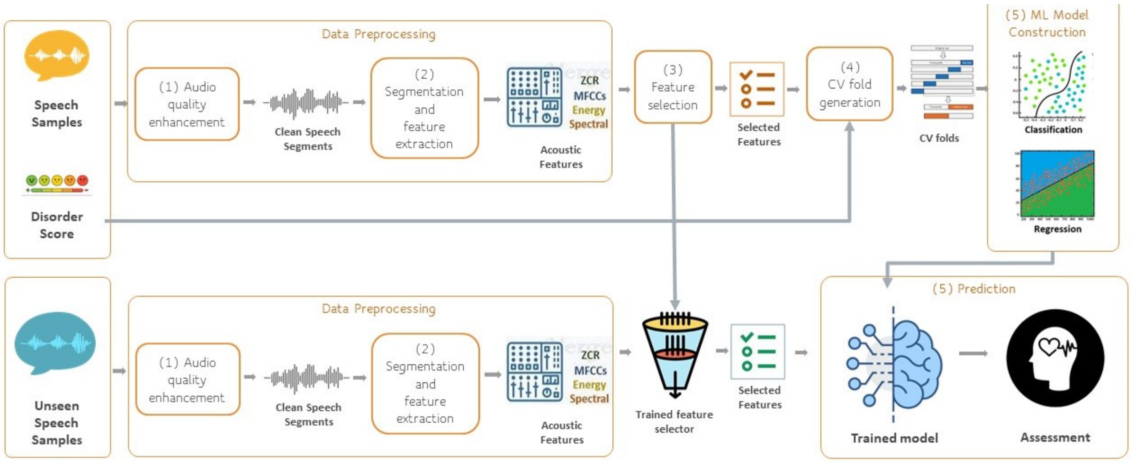


Figure 1.2: Audio analysis pipeline

As shown in the pipeline (Figure 1.2), the training samples undergo an audio enhancement process to ensure good quality samples for training models. If the presence of background noise is noticed through manual inspection, noise suppression is applied. Then, acoustic features are extracted from the audio segments of the speaker separated by pauses longer than the threshold (20 seconds). The features are used for training machine learning models. Once the models have been trained and validated, they are applied to make predictions on previously unseen spoken utterances, which are processed through the same cleaning and feature-extraction steps.

The effectiveness of several algorithms has been examined for predicting the prevalence and severity of psychological disorders like depression, anxiety, and stress, and neurodegenerative diseases such as AD and MCI. Some models are trained to perform

binary classification, to distinguish between patients and healthy individuals, while others perform **regression**, to determine the disorder severity on the same standard scales on which the training data is labeled.

A combination of probabilistic, non-linear, and ensemble models have been studied. These models were found effective for similar tasks in existing literature as listed in Table 1.3

Table 1.3: Machine learning models for predicting mental health condition

Probabilistic model	Hidden Markov model (HMM)
Non-linear model	Logistic Regression (LR) Decision Tree (DT) Support vector machine (SVM) Neural network (NN)
Ensemble model	Random forest (RF) Gradient boosting tree (GBT) Extreme gradient booting (XGBoost)

1.4 Contributions

Experimenting on an array of datasets encompassing psychiatric disorders, this thesis endeavors to reveal the inherent traits in vocal tones to get insight into an individual's mental health and cognitive competency. The thesis makes two key contributions:

Exploration of a variety of datasets: In this thesis we experimented on datasets involving psychiatric disorders, i.e., depression, anxiety and stress, and neurodegenerative disorders like Alzheimer's dementia, and MCI.

Methodical analysis of these datasets sheds light on their Strengths and Weaknesses in gaining important insights into the corresponding disorder.

- Longer, spontaneous speech is more suitable for training models to predict depression than shorter, guided speech task (Chapter 2)

- Multiple speech samples collected in the same recording session do not provide additional information than a single sample (Chapter 4), however, multiple samples collected in separate sessions from an individual provide better prediction of mental health conditions, as described in Chapter 5

Development and validation of a speech processing pipeline: A complete pipeline for speech analysis to measure the prevalence and severity of mental health conditions and cognitive impairment, comprising three major components (data preprocessing, model training, and prediction) has been established through meticulous experimentation (Figure 1.2). The pipeline has been refined by testing on multiple mental health conditions including depression, anxiety, stress, MCI, and AD; speech samples in different languages, e.g., English, German, Spanish, and Chinese, in a variety of recording conditions. Diverse datasets, spanning languages, populations, and conditions, validate its generality. Consistently competitive performance in classification and regression tasks validates its efficacy.

These contributions collectively advance speech-based mental health assessment, promising more effective and accessible diagnosis and intervention methods. Different adaptations of the pipeline, along with the comparison of performance with the state-of-the-art on different tasks have been elaborated in Chapter 2, 3, 4, 5, 6, and 7.

1.5 Organization

The research conducted for this thesis is organized into two broad themes. The analysis techniques and experimental outcomes of several depression-related datasets are discussed in Chapters 2, 3, 4, and 5. The application of a similar methodology in predicting the severity of AD and MCI is discussed in Chapters 6 and 7.

Each chapter initiates with an overview of the detailed experiment, outlining the three components of the speech analysis pipeline and summarizing the findings of the corresponding experiment. The original publications are also included in the chapters for reference. The last chapter of the thesis, Chapter 8, concludes with an overview of the major findings of this work and lays out some promising directions for future research.

Chapter 2

Detecting Depression From Voice

In this chapter, I present the preliminary version of the speech-processing pipeline (Figure 1.2) to analyze the two most popular benchmark depression datasets: AVEC 2013 and AVEC 2017. Both classification of samples from depressed vs. healthy individuals and regression prediction of depression severity were conducted through the following implementation of the pipeline:

- **Data preprocessing:** Samples of both datasets were recorded in controlled environments using high-quality recording equipment, therefore the audio is of sufficient quality. The AVEC 2017 dataset contains speech segments of interviewers alongside the subjects. The interviewers' segments were removed using timestamps provided in the transcripts. 2,268-dimensional AVEC 2013 baseline features are extracted from a) the entire length of audio from the AVEC 2013 dataset and b) each speech segment of the audio samples of the AVEC 2017 dataset. The features were standardized and reduced to 791 dimensions using principal component analysis (PCA) retaining 95% variance.

- **Model training:** Following the literature, SVM, GBT, random forest, and neural network models were trained for each of the classification and regression tasks on the training samples of the datasets.

- **Prediction:** Both datasets included held-out test sets to evaluate the models. The models trained through internal 5-fold CV made predictions on the test partitions. Majority voting and arithmetic mean were calculated to aggregate the classification and regression predictions respectively on each sample's segments. On the test partition of both datasets, the neural network model yielded the highest classification accuracy and the random forest model obtained the lowest regression error.

Key findings: The analysis presented in this chapter showed that:

1. The **semi-structured interview** of the AVEC 2017 dataset provided the **classifiers** with more detailed information compared to the guided speech task samples from the AVEC 2013 dataset. This additional detail likely contributed to the higher accuracy observed in the AVEC 2017 test set.

2. However, a similar phenomenon was not observed in the case of the regression task. AVEC 2013 and AVEC 2017 datasets measure depression severity on BDI-II (range 0-63) and PHQ-8 (range 0-24), respectively. The normalized RMSE across the range of the scales achieved by random forest models were 0.155 and 0.257 respectively on the AVEC 2013 and AVEC 2017 test sets.

   This could be caused by the variation in the formulation of the depression scales, the bias in the dataset, and the method of aggregating the predictions, which was investigated through further experimentation in the later chapters.

This work was published in the Advances in Artificial Intelligence: 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019, Kingston, ON, Canada, May 28–31, 2019.

As the first author, I designed and conducted the experiments, under the supervision of Prof. Eleni Stroulia.

Detecting Depression From Voice

***Abstract*:** In this paper, we present our exploration of different machine-learning algorithms for detecting depression by analyzing the acoustic features of a person's voice. We have conducted our study on benchmark datasets, in order to identify the best framework for the task, in anticipation of deploying it in a future application.

2.1 Introduction

Depression is the most common psychological disorder, affecting more than 300 million people worldwide and is considered the leading cause of disability worldwide [12]. Current depression diagnostic instruments require the active participation of depressed individuals. However, due to lack of awareness and the nature of the disorder itself, a large percentage of the population refrains from seeking expert assistance. Recent studies reveal that depression is reflected in behavioural fluctuations of certain day-to-day activities and also in the ways people talk [13]. These findings have motivated a wave of research efforts aimed at developing automated depression detection methods based on vocal acoustic features. Introduction of Depression Recognition Sub-Challenge (DSC) as a part of the Audio/Visual Emotion Challenge (AVEC) since 2013 has accelerated interventions in depression recognition combining different modalities, i.e., audio, video, and text features [14]. Different directions of feature engineering, algorithms, and contextual information incorporation have been explored in four challenges taken place this far.

In our work, we are interested in developing a practical system that can capture the audio of the users' voices during phone call conversations and analyze it to detect their depression levels. A prerequisite for such a system is a model capable of detecting evidence of depression from conversational audio. In this work, we explored the effectiveness of different machine-learning algorithms for the anticipated depression detection model with the currently available AVEC data sets.

The rest of the paper is organized as follows. Section 2.2 reviews the related research on depression detection based on a subject's vocal biomarkers. We explain the analysis methodology and experimental results in Section 2.3. The paper has been concluded discussing future prospects in Section 2.4.

2.2 Background and Related Work

Much of the work in this area (including ours) has been done on two publicly available benchmark audio datasets: AVEC 2013 and AVEC 2017. The AVEC 2013 audio corpus is a subset of the audio-visual depressive language corpus (AVDLC) [14], where 292 subjects performed two PowerPoint-guided tasks in German while being recorded, resulting in 300 recordings. There is only one person in every recording, and the speakers were recorded between one and four times, with a period of two weeks between the measurements. Each training, development, and test partitions consists of 100 recordings. The training and development partitions are labelled with depression scores in 21 items BDI-II scale ranging from 0-63, where a score greater than 19 is considered to belong in the *"depressed"* class [15]. The mean BDI-II score of the recordings is $\mu = 15.3$ with standard deviation (SD) = 12.3. The AVEC 2017 DSC dataset consists of 189 audio recordings of clinical interviews [16]. The recordings are labelled with depression scores of the participants in the 8-item PHQ-8 depression inventory ranging from 0-24. A score of 10 or greater is considered as major depression. The average depression severity on the training and development set of the challenge is M = 6.67 (SD = 5.75). The training, development and test data sets contain 107, 35 and 47 audio recordings respectively.

Distinguishing depressed individuals from non-depressed ones is a binary classification task, while determining severity by predicting depression score formulates a regression problem. The binary-classification task has been explored on a variety of datasets in [17], [18], [19], [20], [15] and [21]. Prosodic, spectral, cepstral, glottal, energy related features have been experimented with for this purpose. Moore II *et al.* reported superiority of glottal features over prosodic ones [17], while spectral and energy-related features were found most effective by Lopez-Otero *et al.* [15]. Low *et al.* reported significant gender dependency in classification accuracy using Teager energy operator (TEO) features [19]. Besides these, covariance structure of Gaussian Mixture Model (GMM) of recorded speech was found informative by Cummins *et al.* [20]. To overcome small sample size in high dimensional feature space, Moore II *et al.* adopted one-feature-at-a-time strategy and Sanchez *et al.* used backward elimination. The highest classification accuracy was reported by Moore II *et al.* (95.6%) on recordings from 15 depressed and 18 control subjects using quadratic discriminant analysis [17], followed by SVM achieving 87.0%, 81.3% and 65.8% accuracy reported in [19], [18] and [21] respectively. Linear discriminate analysis (LDA), adopted by Lopez-Otero *et al.*, performed the classification task with 70% accuracy on the AVEC 2013 development dataset.

Besides binary classification, researchers have also endeavoured to determine depression severity by predicting depression score using audio features. AVEC 2013 depression sub-challenge (DSC) dataset has been used in [22], [23], [24] and [25] to predict BDI-II score. In [22] combination of formant and delta-mel-cepstral features were used to train Gaussian staircase regression system. Their subject-based and subject-independent adaptation achieved method achieved root mean square error (RMSE) 8.68 and 7.42 respectively. He & Cao used combination of Median Robust Extended Local Binary Patterns (MRELBP) and AVEC 2013 baseline features (mentioned as hand-crafted features in the literature) with deep-learned features extracted from raw audio and spectrogram images for their proposed deep convolutional neural network (DCNN) architecture [23]. The proposed model obtained RMSE 9.89 and mean absolute error (MAE) 8.19. Özkanca *et al.* compared the performance of their proposed framework using Turkish and German (AVEC 2013) dataset. They applied minimum redundancy maximum relevance (MRMR) feature selection criteria on AVEC 2013 baseline feature set of 2268 features prior to using Support Vector Regressor (SVR)[24]. The best RMSE of 9.42 was reported on this dataset. Morales also applied SVR for comparing depression detection systems on several publicly available depression datasets using prosodic and speech rate related features and documented RMSE = 10.70 (MAE 8.59).

The AVEC 2017 DSC dataset was used in [26], [27], [28], [29] and [30]. In addition to challenge beaseline audio features, Sun *et al.* took text topics into account [26], while Gong & Poellabauer considered a more extended set of features, including audio, video and text features. Yang *et al.* extracted deep learned features from spectrograms and Samareh *et al.* added Delta and Delta-Delta coefficients, mean, median, SD, peak-magnitude to RMS ratio to the set of challenge baseline audio features. On the AVEC 2017 dataset best performance was obtained using Deep Convolutional Neural Network (DCNN) and Deep Neural Network (DNN) based audio visual multi-modal depression recognition framework [28], followed by stochastic gradient descent (SGD) regressor [30], random forest [29] and SVM [27]. RMSE 6.32 (MAE 4.40) and 5.45 (MAE 4.32) were reported by [27] and [29] respectively on the development set using audio features exclusively. The challenge baseline RMSE was set 6.74 (MAE 5.36) and 7.78 (MAE 5.72) for development and test set respectively [16].

Considering the fact that conversational audio provides valuable information to assist depression detection, we plan to develop our audio based depression detection system from phone conversation. Here we analyze performance of different classification and regression model for sensing prevalence and severity of depression with a view to finding the best model for future usage.

2.3 Our Method

We applied four algorithms for the classification and regression tasks using the AVEC 2013 [14] and AVEC 2017 [16] data sets.

Data Pre-processing: The audio recording in these data sets is anywhere between 5 and 50 minutes long. The AVEC 2013 corpus provides features extracted from 20s-long windows (shifting forward at the rate of one second). For the AVEC 2017 dataset, we did segmentation based on subjects' voice activity.

In this work, we experimented on the AVEC 2013 baseline feature set consisting of 2,268 audio features. The feature set comprises 76 low-level descriptors (LLD) features and their statistical, regressional, and local minima/maxima-related functionals. The LLD features include energy and spectral-related, voicing-related, delta coefficients of the energy/spectral features, delta coefficients of the voicing-related LLDs, and voiced/unvoiced durational features. We standardized the features by removing the mean and scaling to unit variance. We applied principal component analysis (PCA) to identify the minimum number of features that are capable of retaining 95% of the variance, resulting in 791 features.

Model Training: We trained the following four algorithms for each of the classification and regression tasks on the processed features from the training partitions of the datasets:

Random Forest: 100 estimator trees with a learning rate of 0.1 for both the classification and regression tasks.

Support Vector Machine (SVM): We used the radial basis function (RBF) kernel for SVM.

Gradient Boosting Tree (GBT): For both the GBT classifier and regressor, 100 estimator trees were used.

Deep Neural Network: The network consists of three fully connected layers with

512, 256 and 512 neurons respectively. To avoid overfitting 30-50% dropout was added between layers. We trained the model with mini-batch gradient descent with a batch size of 64. Categorical cross-entropy and mean squared error were considered as the loss function for classification and regression respectively. We exploited the best model chosen from 500 epochs. The learning rate of 10^{-4} was found to best fit for classification while 10^{-3} did well for regression. The Adam optimizer was used for model optimization.

2.3.1 Experimental Results and Discussion

The first question of interest in this study is *"how effective are the chosen algorithms in distinguishing between depressed and non-depressed individuals?"*. The accuracy, precision, and recall of binary classification are summarized in Table 2.1.

Table 2.1: "Depressed" and "Not-depressed" classification

Algorithm	AVEC 2013			AVEC 2017 Dev			AVEC 2017 Test		
	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.
SVM	67.78	0.64	0.68	57.92	0.50	0.58	63.81	0.58	0.64
Random Forest	64.50	0.51	0.64	60.19	0.52	0.60	69.24	0.58	0.69
GBT	62.26	0.57	0.62	58.40	0.54	0.58	63.58	0.59	0.64
DNN	**72.85**	0.70	0.72	**74.65**	0.49	0.56	**80.11**	0.59	0.64

Table 2.2: "Level of depression" Regression

Algorithm	AVEC 2013		AVEC 2017 Dev		AVEC 2017 Test	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
SVM	10.55	**7.93**	7.50	6.11	6.44	5.37
Random Forest	**9.75**	8.21	**6.60**	**5.55**	**6.17**	**5.22**
GBT	14.60	10.38	6.63	5.49	6.26	5.26
DNN	10.74	8.75	8.07	6.67	6.55	5.33
Baseline [14], [16]	11.52	8.93	6.74	5.36	7.78	5.72

Based on Table 2.1, one can see that the deep neural network (DNN) performed best on both data sets. The DNN accuracy for the AVEC 2013 dataset is 72.85%, which is a marginal improvement from the accuracy reported by Lopez-Otero *et al.* [15].

The DNN accuracy is higher with the AVEC 2017 data set, which may likely be attributed to the fact that the AVEC 2013 data set is smaller and may not be sufficient to train the DNN. The DNN achieved 74.65% and 80.11% accuracy respectively on the development and test partition of the AVEC 2017 dataset. The low recall values (0.56 on the development set and 0.64 on the test set) indicate that a significant portion of depressed cases have been misclassified. This is a very undesirable phenomenon when considering applying this method to support real-world diagnosis. One possible reason for this outcome is the imbalanced proportion of class samples in the training set (30 depressed vs 77 not depressed). In the future, we plan to solve this issue by applying a synthetic over-sampling technique to the minority class.

The second question of interest is *"How effective are the chosen algorithms in assessing an individual's level of depression?"*. Results of the regression task are summarized in Table 2.2. The random forest algorithm performed best on both datasets, outperforming the baseline models.

It is important to note, however, that DCNN reported in [23] performs marginally better than our random forest regressor on the AVEC 2013 dataset (RMSE 8.89). As the classes on the BDI-II depression scale are 5 to 8 points apart($<$ 14: minimal, 14-19: mild, 20-28: moderate, $>$28: severe), the current results indicate that there is a high possibility of misclassification, implying room for further improvement.

For the AVEC 2017 dataset, most of the existing models use additional data beyond the voice audio, i.e., including video and text features into their process. As the motivation of our work is to find a reliable model to detect depression prevalence and severity from phone-call conversations, we only took audio features into account. In a person-invariant unimodal (audio only) scenario, results of our random forest model are consistent with results reported in [27] and [29] on the development set. For the test set the only unimodal result is available in the challenge baseline [16]

(RMSE = 7.78, MAE = 5.72) which is outperformed by our proposed model (RMSE = 6.17, MAE = 5.22). Still, the range of scores on the PHQ-8 scale is 0-24 where a score higher than 9 indicates major depression, therefore a more accurate model will increase the reliability of our envisioned system.

2.4 Conclusions and Future Work

In this work, detecting the prevalence and severity of depression from acoustic features of conversational speech in two languages has been explored using different classification and regression algorithms. We have found that the deep neural network performs best in binary classification, while random forest gives competitive results for the regression task. In the future, we will consider synthetically balancing the classes as our next measure for performance improvement.

Chapter 3

Detecting Depression from Speech: A Systematic Exploration of Machine Learners

This chapter describes the revised speech-processing pipeline initiated in Chapter 2. This experiment was conducted using AVEC 2013 and AVEC 2019 (an extended version of AVEC 2017 dataset) datasets. The following speech-processing pipeline was implemented in this experiment:

- **Data preprocessing:** Both the AVEC datasets included high-quality speech samples, therefore audio enhancement was not required in this experiment. A novel data preprocessing step was introduced to aggregate the predictions per segment to achieve the final prediction on the whole speech sample. In this feature-vector construction approach, each feature's value range from all training utterance windows was divided into 30 equally spaced bins. For values beyond the range of training data, -1000 and +1000 are added at the beginning and end of the partitions. The number of bins (here, 30) was selected empirically in this experiment, but there was room for further investigation to determine the optimal number. During the data loading, of each sample, the relative distribution of values from all the segments for each feature in the universal range is determined by calculating the percentage of values falling in each of the 30 bins. (Algorithm 1).

Three different sets of acoustic features were extracted from each audio recording provided in each of the two datasets: 1) AVEC 2013 baseline, 2) INTERSPEECH 2013 ComParE, and 3) eGeMAPS. Usage of these features is frequently found in the literature on depression detection and speech-emotion analysis. The three feature sets include 76, 130, and 23 low-level descriptors (LLDs) respectively, encompassing energy, spectral, MFCC, and voicing-related features. Feature value binning was applied to the LLDs extracted from 25 ms long segments. Statistical functionals were computed on the LLDs to obtain 2,268, 6,373, and 88-dimensional feature vectors for each speech utterance (spoken sentences separated by pauses).

- **Model training:** In the training phase, decision tree, gradient boosting tree, extreme gradient boosting, random forest, and support vector machine models were trained for both classification and regression tasks. The Neural network and hidden Markov models were trained for the classification task only. The machine learning algorithms were selected based on recent literature and expert opinion. Each model was trained separately on feature vectors with and without incorporating *"feature value binning"* to compare the variation in performance offered by the approach.

- **Predictions:** The trained models made predictions on the held-out test sets of both datasets.

Key findings:

1. For the classification task, the AVEC 2013 feature set was found most effective for training DNNs and HMMs, respectively, on both datasets. For the regression task also, the lowest RMSE on the AVEC 2013 test set was achieved by the SVM model using this feature set

Algorithm 1 Pseudo-code for Feature-Vector Construction with Feature-value Binning

```
function CreateBins (Dataframe $df$ containing $n$ features extracted from each of $m$ audio samples)
    $r \leftarrow$ number of rows in $df$
    for all feature $f$ indexing from 0 to $n$ do
        $min \leftarrow$ minimum value of feature $f$
        $max \leftarrow$ maximum value of feature $f$
        $partitions \leftarrow$ 28 equally spaced bin partitions from $min$ to $max$
        $start \leftarrow$ -infinity
        $end \leftarrow$ infinity
        $featureBins \leftarrow$ concatenate($start$, $partitions$, $end$)
        append $featureBins$ to $FeatureBinVector$
    end for
    return $FeatureBinVector$
end function

function featureVectorConstruction($m$ feature files each with $n$ features)
    for all file indexing from 0 to $m$ do
        $r \leftarrow$ number of rows in the file
        for all feature $f$ indexing from 0 to $n$ do
            $bins \leftarrow$ bin partitions for feature $f$ in $FeatureBinVector$
            for all bin index $b$ do
                $count \leftarrow$ number of feature value $v$: lower limit of $bins[b] < v \leq$ upper limit of $bins[b]$
                $distribution[b] = \frac{count}{r}$
            end for
            concatenate $distribution$ to $featureValueDistribution$
        end for
        append $featureValueDistribution$ to $FeatureVector$
    end for
    return $FeatureVector$
end function
```

2. On the AVEC 2019 test set, the random forest model scored the lowest RMSE (6.37) using the ComParE feature set. But the improvement is marginal over the AVEC 2013 feature set (RMSE 6.43) and is out-weighted by the additional computational cost of constructing the ComParE feature set (6,373 features) in comparison to the AVEC 2013 feature set (2,268 features)

3. For both of the tasks, the application of feature value binning offered a significant reduction in computational complexity, with identical or better performance of the models

4. Feature value binning provided a better performance on the comparatively longer English dataset (AVEC 2019)

5. On the shorter-length German dataset (AVEC 2013) application of feature vector extracted per speech sample performed better

In a real-world context, one would expect to collect longitudinal data more realistic than reading a short book passage; therefore, feature value binning could be recommended to be incorporated into the depression prediction pipeline. Although this experiment was insufficient to conclusively recommend the methodology, it provided a strong signal for more research and comparative analysis on both benchmark and real-world data.

The experimental methodology was submitted to IEEE Access.

As the first author, I conducted the experiments under the supervision of Prof. Eleni Stroulia.

Detecting Depression from Speech: A Systematic Exploration of Machine Learners

Abstract:

Objective: Depression is a mood disorder caused by a combination of genetic, biological, environmental, and psychological factors. Studies have shown that depression is reflected in behavioural fluctuations in day-to-day activities, as well as in speech abnormalities. This finding has motivated a substantial body of machine learning research, including ours, to construct models capable of recognizing the prevalence and severity of depression from vocal acoustic biomarkers.

Methods: In this paper, we present our systematic exploration of different data-processing methods and machine learning algorithms for detecting depression by analyzing various acoustic features from speech extracted from publicly available audio datasets; to identify the best combination of audio processing methods and machine learning model for the task.

Result: In our experiments, we have demonstrated that the machine learning models accurately classify 70 to 80% samples of the AVEC 2013 (German) and 2019 (English) datasets, respectively. Our proposed Feature-value Binning preprocessing method significantly reduces the computational complexity of model training, with noteworthy improvement in the performance of the regressors and competitive classification performance.

Significance: The proposed speech-processing and machine learning methodology

can potentially serve as a useful instrument for timely diagnosis and continuous monitoring of depression.

3.1 Introduction

Depression is a common psychological phenomenon. People with depression may experience a lack of interest and pleasure in daily activities, significant weight loss or gain, insomnia or excessive sleeping, lack of energy, inability to concentrate, feelings of worthlessness or guilt and recurrent thoughts of death or suicide [31]. About 264 million people of all ages suffer from depression, which is approximately 5% of the world's total population. Depression is a leading cause of disability and suicide worldwide [12]. Suicide causes one death every 40 seconds, and it is statistically the second leading cause of death among youths between 15–29 years of age [32] worldwide. In addition, it is estimated that 10 to 13 percent of women experience depression during pregnancy or after giving birth, limiting their capacity to provide childcare and ultimately resulting in poor growth and development of their children [33].

Today, the diagnosis of depression relies on a significant amount of time and active participation of the depressed individuals. However, due to the lack of a sufficient number of mental health professionals and the nature of the disorder itself, a large percentage of the population refrains from seeking expert assistance. Studies reveal that depression is reflected in behavioural fluctuations of certain day-to-day activities and is also demonstrated in prosodic speech abnormalities [13], [34]. These findings have motivated a wave of research efforts aimed at developing automated depression detection methods based on vocal acoustic features. The introduction of Depression Recognition Sub-Challenge (DSC) as a part of the Audio/Visual Emotion Challenge (AVEC) has, since 2013, accelerated interventions in depression recognition combining different modalities, i.e., audio, video, and text features [14]. Different directions of (1) contextual data fusion, (2) feature engineering and (3) machine learning algorithms have been explored in six challenges taken place thus far. The three challenge datasets, being publicly available, have paved the way for researchers to enhance the capabilities of depression detection models.

The field of detecting depression (and other mental health conditions) from voice is very active and quite divergent: different papers use different data sets, extract different features, and report different metrics. One of our objectives, and one of the contributions, of our paper is to comparatively investigate different combinations of acoustic features and machine-learning algorithms for this task. Although this area has been widely investigated from different perspectives, the task is still challenging and deserves our continued attention. The long-term objective of our work is to develop a system that will detect depression by analyzing the sound of users' voices. Such a system will potentially be helpful for healthcare professionals, such as psychologists and psychiatrists, in the continuous monitoring of their patients with the use of sensors embedded in everyday smart devices, like smartphones, and to help them provide timely support for depressed individuals. A prerequisite for our envisioned system is a scalable and computationally efficient model, capable of detecting evidence of depression from conversational audio. In this work, we have explored several feature sets used in the literature and proposed a novel data processing methodology and feature aggregation method. We addressed the following research questions:

- How informative is audio as a modality to detect the prevalence (classification) and severity (regression) of depression?
- What is the appropriate feature set, effectively balancing output quality and computational cost?
- What is the most effective (type of) algorithm and configuration for the above two tasks?

We have explored the effectiveness of different preprocessing methods and machine-learning algorithms to construct the depression detection model with two of the publicly available AVEC data sets.

The rest of this chapter is organized as follows. Section 3.2 reviews the related research on depression detection based on a subject's vocal biomarkers. Analysis of the characteristics of the two data sets used in this work is also presented in this section. We explain the analysis methodology and experimental setup in Section 3.3 and 3.4 respectively. Section 3.5 presents the performance evaluation of the proposed methodologies. The chapter has been concluded, discussing prospects in Section 3.6.

3.2 Background and Related Research

Studies show evidence of measurable fluctuation in vocal parameters among individuals suffering from psychological and neurological disorders, including depression [35], [36], cerebral palsy [37], amyotrophic lateral sclerosis [38], Parkinson's disease [39] etc. In this section, the outcomes of research in the area of audio-based depression detection are summarized.

3.2.1 Audio Corpora for Depression-Detection Research

A significant number of studies have been conducted to sense the psychological state of individuals based on speech characteristics; a considerable portion of which has been done on one or more of the three publicly available AVEC DSC datasets: AVEC 2013, AVEC 2017, and AVEC 2019. As the AVEC 2019 is a superset of the AVEC 2017 dataset, we conducted our analysis on the AVEC 2013 and AVEC 2019 datasets.

AVEC 2013 DSC Dataset

The dataset [14] consists of recordings from 84 subjects, aged between 18 and 63 years (M = 31.5, SD = 12.3). Every participant performs two tasks while being recorded: (a) reading aloud a part of the fable "The North Wind and the Sun", and (b) answering a question such as "What is your favorite dish?"; "What was your best gift, and why?"; or "Discuss a sad childhood memory", both in German. The recordings are divided into three partitions: a training, development, and test set of 150 Northwind-Freeform pairs. The depression labels are available only for the training and development partitions, therefore in our work, we could use only these two partitions. The duration of the recordings ranges from 6 to 248 seconds.

The recordings are labeled with the Beck Depression Index (BDI) of the speakers. The scores range from 0 to 63, where 0 indicates minimal depression and 63 indicates most severe depression. The highest BDI score in this dataset is 45, which may affect the performance of the proposed system in real-world scenarios.

For the classification task we consider BDI scores ranging from 0 to 19, corresponding to minimal and mild depression levels, as "non-depressed" samples; scores higher than 19, corresponding to moderate and severe depression levels, are considered as "depressed" samples [15]. The partitions are not balanced in terms of depression severity; the ratio of "depressed" and "non-depressed" samples are 1:2 and 2:3 in the training and development set, respectively.

AVEC 2019 DSC Dataset

This dataset is an extended version of the DAIC-WoZ [40], consisting of semi-clinical interviews. The interviews were conducted to create a computer agent capable of interviewing people and identifying verbal and nonverbal indicators of mental illnesses [41]. In the WoZ interviews, the virtual agent is controlled by a human interviewer (wizard) in another room, whereas in the AI interviews, the agent acts in a fully autonomous way. Participants were recruited through two channels: online ads posted on Craigslist.org, and on-site at a US Vets facility in Southern California [40]. Transcribed text of the recordings using Google Cloud's speech recognition service is also included in the dataset.

The audio recordings are labelled with self-reported eight-item Patient Health Questionnaire (PHQ-8) scores between 0 and 24; samples with scores higher than 12 are considered as "depressed" and the rest as "non-depressed". The dataset is partitioned into training, development, and test sets containing 163, 56 and 56 recordings respectively. Although there is at least one sample labelled with each of the 24 PHQ-8 scores in the training set, the fact does not hold for most of the PHQ-8 scores higher than the "depressed" class threshold in the development set. The number of samples with low PHQ-8 is significantly higher than the number of samples with high PHQ-8 scores. In fact, the number of "non-depressed" samples is about 3 times higher than the number of "depressed" samples in the training and development partitions. The training and development sets include a combination of WoZ and AI scenarios, the test set only contains data collected by the autonomous AI. The length of the WoZ interviews ranges from 5 to 20 minutes, and the automated ones from 15 to 25 minutes.

3.2.2 Performance Evaluation Metrics

Classification

The performance of depression-classification methods is typically evaluated in terms of accuracy and F_1 score.

$$Accuracy = \frac{TP + TN}{N} \tag{3.1}$$

and

$$F_1 = 2 \times \frac{\pi \times \rho}{\pi + \rho} \tag{3.2}$$

where

$$precision\ \pi = \frac{TP}{TP + FP} \tag{3.3}$$

and

$$recall\ \rho = \frac{TP}{TP + FN} \tag{3.4}$$

In the above equations, N is the number of samples; TP is the number of samples correctly classified as positive; TN is the number of samples correctly classified as negative; FP is the number of samples incorrectly classified as positive; and FN is the number of samples incorrectly classified as negative.

Regression

Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are the most commonly used performance metrics of regressors. In addition, following AVEC 2019 Detecting Depression with AI Sub-challenge (DDS) specifications, we have evaluated the performance of the regressors using *Concordance Correlation Coefficient (CCC).* The corresponding formulas are shown below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \tag{3.5}$$

$$MAE = \frac{\sum_{i=1}^{N}|x_i - y_i|}{N} \tag{3.6}$$

In the above, x_i and y_i are the true and predicted scores respectively.

CCC measures the correlation between two variables, e.g., ground truth depression score and the predicted score. Its value range is [-1,1], where 1 indicates perfect agreement between the true and the predicted scores.

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{{\sigma_x}^2 + {\sigma_y}^2 + (\mu_x - \mu_y)^2} \tag{3.7}$$

where μ_x and μ_y are the means for the true and predicted score and σ_x^2 and σ_y^2 are the corresponding variances. ρ is the correlation coefficient between the two variables, calculated as:

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{N}(y_i - \mu_y)^2}} \tag{3.8}$$

3.2.3 Audio-Based Depression Detection

Distinguishing depressed individuals from non-depressed ones is a binary classification task, while determining severity by predicting depression score formulates a regression problem. Numerous experiments have been carried out to solve both of them, exploring a wide variety of vocal parameters. We have summarized them in two categories: works on the AVEC DSC datasets and works on other datasets. A summary of experimental results presented in this section using audio modality is presented in Table 3.1 and 3.2.

Table 3.1: Classification results of audio-based depression detection experiments

Dataset	**Literature**	**Accuracy (%)**
15 depressed and 18 control subjects	Moore II *et al.*, 2008 [17]	89.7 (male) 95.6 (female)
16 depressed and 16 control elderly males	Sanchez *et al.*, 2011 [18]	81.3
68 clinically depressed and 71 control adolescents	Low *et al.*, 2011 [19]	87.0 (male) 79.0 (female)
17 depressed and 14 remitted females	Laosaphan and Yingthawornsuk, 2012 [42]	44.0
Mundt database [43]	Cummins *et al.*, 2013 [20]	69.0
Black Dog database [44]	Cummins *et al.*, 2013 [20]	63.0
The AVEC 2013 DSC Dataset	Lopez-Otero *et al.*, 2014 [15]	70.0
Pitt corpus in the DementiaBank database [45]	Fraser *et al.*, 2016 [21]	65.8
SH2 dataset	Huang *et al.*, 2017 [46]	72.9
CONVERGE dataset	Afshan *et al.*, 2018 [47]	94.79
DAIC-WoZ Dataset [40]	Salekin *et al.*, 2018 [48]	96.7
	Muzammel *et al.*, 2020 [49]	86.06
Mundt database [43]	Seneviratne *et al.*, 2020 [50]	81.77

Research on the AVEC DSC Datasets

Lopez-Otero *et al.* presented the accuracy of the Gaussian Mixture Model in classifying speech as “depressed” or “non-depressed” in the AVEC 2013 DSC framework [15]. Cepstral, prosodic, spectral, and energy features were used, spectral and energy-related features obtained the highest accuracy of 70% in these experiments irrespective of applying linear discriminant analysis (LDA). The authors discussed that the limitation of the small database size and particularly insufficient number of examples of the different depression levels poses uncertainty in whether their proposed depression level classifier is generalizable to other data.

Salekin *et al.* introduced weakly supervised NN2Vec feature modelling with a fully connected neural network to construct feature vectors using energy and spectral low-level descriptors (LLDs) and statistical functionals [48]. Their multiple instance learning (MIL) adaptation of the BLSTM classifier achieved 96.7% classification accuracy on DAIC-WoZ [40] corpus in leave-one-out cross-validation. The same corpus was also used by Dubagunta *et al.*, who proposed a convolutional neural network (CNN) architecture for depression recognition from voice source-related features including low pass filtered (LPF), the linear prediction residual (LPR), homomorphically filtered voice source (HFVS), zero frequency filtered (ZFF) signals [63]. On the classification task, an overall F_1 score of 0.69 was achieved on the test partition of the dataset (AVEC 2016 DSC) using subsegmental level modelling of linear prediction residual signals or zero frequency filtered signals.

Muzammel *et al.* discussed the effectiveness of audio vowels and consonants spectrogram-based deep learning descriptors for depression classification. Their proposed deep convolutional neural network (DCNN) architecture achieved 86.06% accuracy on randomly selected 10% test samples of DAIC-WoZ [40] dataset. The authors commented that deep-learned consonant-based acoustic characteristics lead to better recognition results than vowel-based ones [49].

While the above-mentioned studies were focused on binary classification, researchers have also endeavoured to predict depression severity by analyzing audio features. The AVEC 2013 dataset has been used in [22], [23], [24] and [25] to predict BDI-II score [64]. In [22] combination of formant and delta-mel-cepstral features extracted from audio segments of paragraph reading task were used by Williamson *et al.* to train the Gaussian staircase regression system. Their subject-based and subject-independent adaptation achieved RMSE 8.68 and 7.42 respectively. Jan *et al.* introduced the feature dynamic history histogram (FDHH) method to construct a feature vector from the challenge baseline feature set [51]. They experimented on partial least square (PLS) and linear regression (LR) models, and reported RMSE 10.08 and MAE 8.25. He & Cao used a combination of Median Robust Extended Local Binary Patterns

(MRELBP) and AVEC 2013 baseline features (mentioned as hand-crafted features in the literature) with deep-learned features extracted from raw audio and spectrogram images for their proposed DCNN architecture [23]. The proposed model obtained RMSE 9.89 and MAE 8.19.

Özkanca *et al.* mentioned the challenges posed by cultural and linguistic variation in the expression of depression, as they compared the performance of their proposed method using Turkish and German (The AVEC 2013) dataset. They applied minimum redundancy maximum relevance (MRMR) feature selection criteria on the AVEC 2013 baseline feature set (2268 features) prior to training Support Vector Regressor (SVR). The best RMSE of 9.42 was reported on this dataset [24]. Morales also applied SVR to compare depression detection systems on several publicly available depression datasets in unimodal and multimodal (audio, visual, and text mode) format. For audio-based depression detection, prosodic and speech rate-related features were reported to result in RMSE 10.70 (MAE 8.59). In their recent work, Zhao *et al.* discussed the fusion of Self-attention network and DCNN on a combination of eGeMAPS and 3D log-meals spectrogram extracted in a window size of 25 ms and a 10 ms stride [52]. Their proposed technique could obtain RMSE and MAE of 9.65 and 7.38 respectively on this dataset.

The AVEC 2017 DSC dataset is the most widely used depression corpus in the existing literature as its recordings follow clinical interview protocols, are recorded in English, and have longer samples than the AVEC 2013 dataset. To date, researchers are experimenting with a wide spectrum of acoustic features and machine-learning models using this dataset. Williamson *et al.* extracted formants, Mel Frequency Cepstral Coefficients (MFCCs), glottal features, and loudness [53]. In addition to challenge baseline COVAREP [65] audio features, Sun *et al.* took text topics into account [26], while Gong & Poellabauer considered a more extended set of features of audio, video, and text modalities [30]. On the other hand, Yang *et al.* extracted deep-learned features from spectrograms in [28]. Samareh *et al.* added Delta and

Delta-Delta coefficients, mean, median, standard deviation, and peak-magnitude to RMS ratio to the set of challenge baseline audio features [29]. Applying similar higher-order statistics on the baseline features, Alhanai *et al.* constructed an extended feature set of 553 features, of which they identified 279 features with statistically significant univariate correlation [54]. Haque *et al.* implemented multi-modal sentence-level embedding on the log-Mel spectrogram and MFCC features [55]. In their work, Yang *et al.* exploited a combination of eGeMAPS and INTERSPEECH features extracted from the longest ten segments of each audio sample. They reshaped the feature vector in an image-like 2D feature map in row-major order and adopted Deep Convolutional Generative Adversarial Net (DCGAN) for feature vector generation [56].

On the development dataset, Yang *et al.* obtained the best RMSE (3.09) by using a DCNN and Deep Neural Network (DNN) based audiovisual multi-modal depression recognition framework [28], followed by stochastic gradient descent (SGD) regressor (3.54) adopted by Gong & Poellabauer [30], Random Forest (5.45) used by Samareh *et al.* [29], DCNN (5.52) constructed by Yang *et al.* [56], Casual CNN (5.78) proposed by Haque *et al.* [55] and Support Vector Machine (SVM) (6.32) reported by Dham *et al.* [27]. Williamson *et al.* obtained RMSE 6.38 applying Gaussian staircase model [53], while RMSE 6.5 was reported by both Syed *et al.* [57] and Alhanai *et al.* [54] by using partial least square regressor (PLSR) and long short-term memory (LSTM) respectively. RMSE 4.99, 5.40, and 6.42 were reported by Gong & Poellabauer, Yang *et al.* and Syed *et al.* respectively on the test set. The challenge baseline RMSE 6.74 (MAE 5.36) and 7.78 (MAE 5.72) were set for the development and test set respectively [16].

By the time of this thesis, relatively fewer studies have focused on the AVEC 2019 DDS dataset, which is a superset of the AVEC 2017 dataset. Most of the challenge participants have adopted multimodal methods involving acoustic, visual, and linguistic features [66], [58], [60], [59] and [61]. A wide range of audio features has been provided as the challenge baseline including extended Geneva Minimalistic

Acoustic Parameter Set (eGeMAPS), Mel Frequency Cepstral Coefficients (MFCCs), Bag of Audio Words (BoAW) and two sets of deep spectrum features by feeding spectral images of speech instances into pre-trained image recognition Convolutional Neural Networks (CNN) (VGG-16 [67] and DenseNet-121 [46]) and extracting the resulting activations as feature vectors. Of these, deep spectrum features were used by Yin *et al.* [66] and Rodrigues *et al.* [59], and MFCCs and eGeMAPS by Fan *et al.* [61]. Ray *et al.* exploited all the baseline feature sets, while Zhang *et al.* extracted the AVEC 2017 baseline feature set [16] using the COVAREP software toolbox [65], in addition to eGeMAPS. Different configurations of long short-term memory (LSTM) networks were used in all of these works except Zhang *et al.* [60], who adopted random forest and logistic regression. Concordance Correlation Coefficient (CCC) was reported in these works in addition to RMSE following the challenge framework.

Works on Other Datasets

A study conducted by Stefan *et al.* [68] suggests that depressed patients often display a significant reduction in vowel space. In this study, the authors exploited k-means clustering in two-dimensional space of the first two formants extracted from voiced segments to assess a speaker's vowel space. They evaluated their hypothesis on recordings of 253 individuals and found a significantly reduced vowel space in subjects that scored positively on the PHQ-9 scale for depression assessment.

Fraser *et al.* [21] investigated whether automated screening algorithms for Alzheimer's disease (AD) are affected by depression, and attempted to detect when individuals diagnosed with AD also show signs of depression. Their linguistic analysis achieved 65.8% accuracy in detecting signs of depression in AD patients using an SVM classifier. They argued that age and other late-life health conditions pose extended difficulty in the task of diagnosis.

Researchers have been studying reliable, computationally convenient audio features to detect depression for many years. With a sample data size of 33 subjects (15 patients and 18 control), Moore II *et al.* adopted a feature-selection strategy by adding one feature at a time to find the highest classification accuracy through quadratic discriminant analysis [17]. Based on univariate analysis, the authors concluded that the influence of glottal features was the important discriminating factor in improving the detection of clinical depression (74.7% to 89.7% in the male group, 87.8% to 95.6% in the female group) from exclusive usage of prosodic features. The authors acknowledged the limitation in generalization due to a small sample size and the difficulty in extraction of the glottal features from the acoustic speech signal without an electroglottograph device.

Sanchez *et al.* studied prosodic speech measurements (pitch and energy), in addition to spectral features (formants and spectral tilt), and computed statistics of these features over different regions of the speech signal in detecting severe depression of elderly males [18]. A set of 25 out of 90 initial features selected by backward elimination performed best (accuracy 81.3%) for SVM classifier. However, the results of this work cannot be applied to all age groups.

Laosaphan and Yingthawornsuk [42] discussed MFCC extracted from speech samples of 17 depressed and 14 remitted females in depression detection. They reported that the Maximum Likelihood (ML) classifier obtained the best accuracy of 44% when experimented on 50% split of training and test set. This indicates the requirement to include other acoustic features for the improvement of depression classification accuracy.

Low *et al.* investigated acoustic correlates of depression in a sample of 139 adolescents (68 clinically depressed and 71 controls) during family interaction [19]. A combination of prosodic, cepstral, spectral, glottal features, and Teager energy operator (TEO) features were tested within a binary SVM classification framework. The authors reported significant gender differences in classification accuracy ranging between 81%-87% for males and 72%-79% for females using TEO features. However, authors acknowledged challenges imposed by genetic, psychological, social, cultural, and environmental factors that contribute to the development of depression detection system.

Cummins *et al.* investigated the hypothesis that important depression-based information can be captured within the covariance structure of a Gaussian Mixture Model (GMM) of recorded speech [20]. Their analysis shows that variance-only adaptation either outperforms or matches the standard mean-only adaptation when classifying depression with maximum accuracy of 69% on the Mundt database and 63% on the Black Dog database. The authors argued that their comparatively lower accuracy resulted from different adaptations of GMMs as classification systems.

Huang et al. introduced SH2 dataset containing around 16 hours of real-world speech data collected from 887 participants (436 female and 450 male), recorded using a variety of different smartphones and labelled with PHQ-9 scores [46]. After segmentation through voice activity detection (VAD), they extracted the INTERSPEECH 2010 [69] baseline LLDs to train a linear SVM classifier. Their proposed system yielded 72.9% accuracy through 3-fold cross-validation. The authors remarked that "conservative segment selection strategies using highly thresholded voice activity detection, coupled with tailored normalization approaches are effective for mitigating smartphone channel variability and background environmental noise".

Afshan *et al.* conducted their experiment on the CONVERGE dataset, which includes recordings of the interviews from 735 individuals classified as suffering from MDD and 953 healthy individuals, in Mandarin [47]. They constructed i-vectors from ComParE 2016 [70] baseline, MFCC and voice quality features using universal background Gaussian mixture model (UBM). The classification accuracy of their logistic regression model was 94.79% on randomly selected 30% test samples.

Seneviratne *et al.* applied DNN based acoustic-to-articulatory speech inversion (SI) on the vocal tract variables and MFCCs extracted from Mundt database [50]. Their proposed SVM classifier achieved 81.77% accuracy through leave-one-subject-out (LOSO) cross validation.

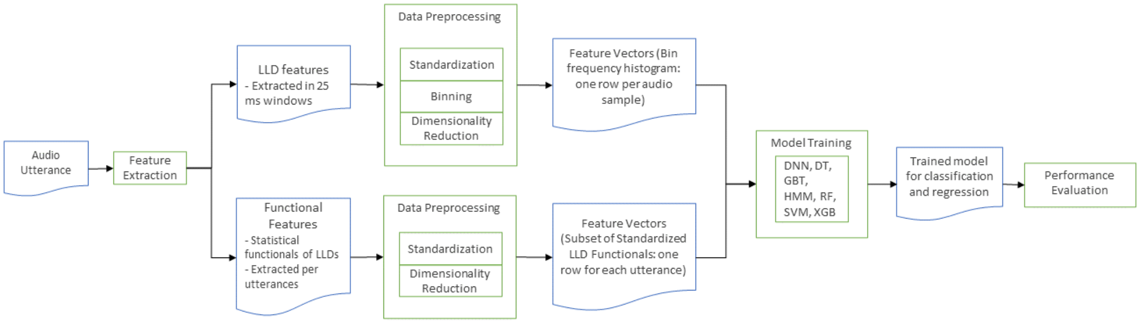

Audio Utterance
Feature Extraction
LLD features
- Extracted in 25 ms windows
Functional Features
- Statistical functionals of LLDs
- Extracted per utterances
Data Preprocessing
Standardization
Binning
Dimensionality Reduction
Data Preprocessing
Standardization
Dimensionality Reduction
Feature Vectors (Bin frequency histogram: one row per audio sample)
Feature Vectors (Subset of Standardized LLD Functionals: one row for each utterance)
Model Training
DNN, DT, GBT, HMM, RF, SVM, XGB
Trained model for classification and regression
Performance Evaluation

Figure 3.1: Pipeline for training machine learning models to predict depression from speech

Considering that conversational audio provides valuable information to assist depression detection, we plan to develop our audio based depression detection system from natural conversation.

3.3 Methodology

We applied the standard processing pipeline shown in Figure 3.1 for analyzing audio data. Each step of the process is discussed in detail below.

3.3.1 Feature Extraction

We extracted three different sets of acoustic features from each audio recording provided in each of the two datasets. Usage of these features is frequently found in the literature on depression detection and speech emotion analysis.

The **AVEC 2013** [14] feature set was the baseline feature set for the AVEC 2013 DSC. Since then the feature set has been effectively used for detecting depression from voice [24], [23], [71], which is why it is an obvious choice for our experiment. This feature set includes 2,268 acoustic features including 76 low-level descriptors (LLD) features and their statistical, regression and local minima/maxima-related functionals. The LLD features include energy, spectral, and voicing-related features; delta coefficients of the energy/spectral features, delta coefficients of the voicing-related LLDs and voiced/unvoiced duration-based features.

The **INTERSPEECH 2013 ComParE** [72] feature set was the baseline feature set for the first Computational Paralinguistics Challenge (ComParE) 2013. The ComParE feature set comprises 130 LLDs including energy, spectral, MFCC, and voicing-related features, logarithmic harmonic-to-noise ratio (HNR), voice quality features, Viterbi smoothing for F0, spectral harmonicity and psycho-acoustic spectral sharpness. Statistical functions are also computed, leading to a total of 6,373 features. The feature set was found effective in different speech emotion recognition tasks [73] including the detection of psychological disorders such as dementia [74], [75].

The **(eGeMAPS)** feature set is composed of 88 features that are computed by applying statistical functionals on the LLDs including pitch, jitter, shimmer, loudness, harmonics to noise ratio (HNR), spectral slope, alpha ratio, Hammarberg index, formant 1–3 frequency and relative level, formant 1-3 bandwidth, harmonic ratios (H1–H2, H1–A3), spectral energy proportions (0–500Hz and 0–1000 Hz), MFCC 1–4, linear pitch and spectral flux [73]. This feature set is the outcome of the effort to formulate a minimalistic feature set for analyzing voice data utilizing knowledge from prior analysis of emotional speech by psychologists, in contrast to the large-scale, brute-forced feature sets such as AVEC 2013 and ComParE.

The gender of the participants provided with the AVEC 2019 dataset is added as an additional feature for this dataset.

3.3.2 Feature-Value Standardization

The range of values of audio features tends to vary widely. To ensure the even contribution of all features in the classification and regression task, and to speed up gradient descent convergence of the deep neural network, we standardized the features using z-scores – i.e., subtracting the mean and dividing by the standard deviation. The standard score of a sample x of feature f_i is calculated as:

$$z = \frac{x - \mu}{\sigma} \tag{3.9}$$

here μ and σ are the empirical mean and standard deviation of the values of f_i in all training samples.

Centering and scaling is performed independently on each feature, by computing the mean and standard deviation on the training set. These same values of the mean and standard deviation are subsequently used on the test data.

3.3.3 Feature-Vector Construction

An important methodological question in audio-based depression modelling is how to compute an overall label (score) for the input audio sample captured from the subject. Majority voting, thresholding, stacking, etc. are some commonly used accumulation techniques to combine utterance-level labels (scores) into an overall label (score). These techniques are useful when most of the utterance windows contain meaningful information, such as is the case with windowing based on voice activity, or sliding windows of one second or longer. These types of windows, however, are not as effective in capturing the subtle fluctuations in human voice reflecting emotional characteristics, which is the reason that segments as short as few milliseconds are preferred as the basis of model learners. Unfortunately, these short segments often do not contain meaningful information, therefore models depending on traditional decision-accumulation techniques may perform poorly in this scenario.

In this work, the LLDs of the three feature sets were extracted in 25 ms long sliding windows shifting forward by 10 ms. This resulted in 5,312 and 13,030,684 utterance windows from the training sets of the AVEC 2013 and the AVEC 2019 datasets respectively.

We experiment with two different approaches towards inferring an overall label (score): (a) a traditional utterance-based aggregation, and (b) aggregation at the level of feature-vector construction.

Utterance-based Feature-Vector Construction: As is typical with most methods in this field, we constructed feature vectors based on the functional features of each of the three feature sets, discussed above, after dimension reduction. At first, Principal Component Analysis (PCA) [76] was performed to identify the most independent features. We chose to select a minimum number of features that is capable of retaining 95% of the variance. The AVEC 2013 is a comparatively smaller dataset, with only 672 utterances in the training partition. PCA on the functional features extracted from this dataset reduced the dimensionality of AVEC 2013, ComParE 2013 and eGeMAPS from 2,268, 6,373 and 88 to 333, 362 and 43 respectively. The recordings of the AVEC 2019 dataset are comparatively longer, with 15,127 utterances in the training set. In the case of this dataset 773, 1,343 and 49 principal components were selected from the three feature sets respectively preserving 95% information.

Next, we selected the best 50 principal components from the AVEC 2013 and ComParE, and the best 30 from eGeMAPS applying a univariate feature selection method based on ANOVA F-value [77] between binary labels and features for the classification task. In the case of the regression task, the same number of principal components were selected using p-values calculated from training-set features and target depression scores.

Sample-based Feature-Vector Construction with Feature-value Binning: In this novel approach to feature-vector construction, we divided the value range (after normalization) of each feature from all training utterance windows into 30 equally spaced bins. For values beyond the range of training data, -1000 and +1000 are added at the beginning and end of the partitions. In our experiment, we selected the number of bins on an ad-hoc basis, but there is room for further investigation to determine the optimal number. During data loading, of each sample, the relative distribution of values from all the segments for each feature in the universal range is determined by calculating the percentage of values falling in each of the 30 bins. For a feature set with n features, each distributed into b bins, this method will result in an array of length $n \times b$ for every sample. Algorithm 2 describes this method in detail. PCA is then performed on the resulting feature vector for dimension reduction, preserving

99% variance.

Algorithm 2 Pseudo-code for Feature-Vector Construction with Feature-value Binning

```
function CreateBins (Dataframe $df$ containing $n$ features extracted from each of $m$ audio samples)
    $r \leftarrow$ number of rows in $df$
    for all feature $f$ indexing from 0 to $n$ do
        $min \leftarrow$ minimum value of feature $f$
        $max \leftarrow$ maximum value of feature $f$
        $partitions \leftarrow$ 28 equally spaced bin partitions from $min$ to $max$
        $start \leftarrow$ -infinity
        $end \leftarrow$ infinity
        $featureBins \leftarrow$ concatenate($start$, $partitions$, $end$)
        append $featureBins$ to $FeatureBinVector$
    end for
    return $FeatureBinVector$
end function

function featureVectorConstruction($m$ feature files each with $n$ features)
    for all file indexing from 0 to $m$ do
        $r \leftarrow$ number of rows in the file
        for all feature $f$ indexing from 0 to $n$ do
            $bins \leftarrow$ bin partitions for feature $f$ in $FeatureBinVector$
            for all bin index $b$ do
                $count \leftarrow$ number of feature value $v$: lower limit of $bins[b] < v \leq$ upper limit of $bins[b]$
                $distribution[b] = \frac{count}{r}$
            end for
            concatenate $distribution$ to $featureValueDistribution$
        end for
        append $featureValueDistribution$ to $FeatureVector$
    end for
    return $FeatureVector$
end function
```

Application of this method keeps the computational complexity of the model training limited to the order of the number of samples, irrespective of the length of the sample, or the number of segments extracted from it. As the bin partitions are preserved for later usage, incremental training of the model is possible upon the availability of new training data.

3.3.4 Model Training

Next, our system fed the feature vectors to various machine-learning algorithms, to identify patterns of features that can distinguish between “depressed” and “non-depressed” subjects (the classification task), and can compute a subject’s depression score on a respective scale (the regression task). We explored several learning algorithms, including RF [78], gradient boosting (GBT) [79], extreme gradient boosting (XGB) [80], SVM [81] and decision tree (DT) [82] for each of the two tasks. We adopted hidden Markov model (HMM) [83] and neural network (NN) [84] for the classification task only.

3.4 Experimental Setup

3.4.1 Data preparation

We extracted the segments with voice activities, i.e. **utterances** of only the participants from the audio recordings using the timestamps and voice activity duration information, discarding long pauses and noise. The participants’ utterances were isolated from the AVEC 2013 dataset, removing noisy and unvoiced segments, using the information on voice activities of the audio files provided with the dataset. The audio recordings of the AVEC 2019 dataset include utterances of both the participants and the animated virtual interviewer called Ellie. The audio transcripts of this dataset provide the start timestamp and duration of the participants’ utterances. We used this information to discard the interviewer’s speech, keeping the segments with only participants’ utterances. The transcription of this dataset is done using Google Cloud’s speech recognition service, therefore each meaningful sentence is often divided into several *utterances* in the transcripts. We normalized the audio volume across all speech segments to -20 dBFS (DeciBels relative to Full Scale).

Then we extracted the acoustic features described in Section 3.3.1 using OpenSMILE v2.1 [85] software toolkit. The LLDs and the functionals are considered separately for each set of features in our experimental setup. Functional features were extracted from each utterance while the LLDs were extracted in 25 ms windows, shifted forward by 10 ms from each utterance. Each of the Freeform-Northwind pairs of the AVEC 2013 datasets is produced by the same participants, hence the pairs are considered as single samples, and features extracted from both recordings are concatenated before proceeding to further analysis. We used the Standard Scaler function from Scikit-learn preprocessing library [86] to normalize the features. Then we applied each feature vector construction method separately on the normalized functional features and LLDs.

3.4.2 Depression Modeling

In our experimental set-up for depression modelling, we constructed the neural network consisting of 3 fully connected layers with the number of neurons ranging from 16 to 1028. We trained the model with mini-batch gradient descent and used the Adam optimizer [87] for model optimization. A hidden Markov Model with Gaussian emissions was also used for classification. The number of states was set to 2, while the number of iterations varied from 50 to 500 based on training data size. For the other models, our learners use grid search with internal 5-fold cross-validation to tune the hyperparameters in possible cases. In the other cases where the search space was too wide to perform a grid search, a random search was carried out. A list of all classifiers and regressors of our experiment and details on their parameter spaces have been provided in Table 3.3.

For the feature vectors constructed using Feature-value Binning, each classifier produced one prediction of binary class ("depressed" or "non-depressed") and each regressor predicted one depression score (BDI-II or PHQ-8) per audio sample; so there is no need of decision aggregation. For Feature Vector Per Utterance, predictions were made in two steps. In the first step, the classifiers and regressors were trained and tested with the features to make one prediction for each utterance of an audio sample. Next, majority vote classification was performed to assign each sample a binary label, based on the majority labels of the utterance level classification. The predicted depression scores of all the segments of one sample were averaged to calculate the final depression score of that sample.

We ran our experiment on an Intel Core i7-8565U CPU at a clock speed of 1.80-1.99 GHz. The 64-bit Windows operating system was installed on the machine. The system availed 16 GB memory. For feature extraction, we used the OpenSMILE software toolkit. The data preprocessing and model training was done in Python programming language.

3.5 Results and Discussion

Table 3.2: Regression results of audio-based depression detection experiments

Dataset	**Literature**	**RMSE**	**MAE**	**CCC**
The AVEC 2013 DSC Dataset	Williamson *et.al.*, 2013 [22]	8.68	7.12	-
	Jan *et al.*, 2017 [51]	10.08	8.25	-
	He and Cao, 2018 [23]	9.89	8.19	-
	Özkanca *et al.*, 2018 [24]	9.42	-	-
	Morales, 2018 [25]	10.70	8.59	-
	Zhao *et al.*, 2020 [52]	9.65	7.38	-
	Valstar *et al.* [14] (challenge baseline)	10.75	8.66	-
The AVEC 2017 DSC Dataset (Development)	Williamson *et al.*, 2016 [53]	6.38	-	-
	Sun *et al.*, 2017 [26]	5.50	4.31	-
	Gong & Poellabauer, 2017 [30]	3.54	2.77	-
	Yang *et al.*, 2017 [28]	3.09	2.48	-
	Samareh *et al.*, 2018 [29]	5.45	4.52	-
	Alhanai *et al.*, 2018 [54]	6.50	-	-
	Haque *et al.*, 2018 [55]	5.78	-	-
	Yang *et al.*, 2020 [56]	5.78	-	-
	Ringeval *et al.*, 2017 [16]	6.74	5.36	-
The AVEC 2017 Dataset (Test)	Gong & Poellabauer, 2017 [30]	4.99	3.96	-
	Yang *et al.*, 2017 [28]	5.40	4.36	-
	Syed *et al.*, 2017 [57]	6.42	-	-
	Ringeval *et al.*, 2017 [16]	7.78	5.72	-
The AVEC 2019 Dataset (Development)	Ray *et al.*, 2019 [58]	5.11	-	-
	Rodrigues *et al.*, 2019 [59]	5.70	-	0.497
	Zhang *et al.*, 2019 [60]	5.83	-	-
	Fan *et al.*, 2019 [61]	6.20		0.348
	Ringeval *et al.*, 2019 [62] (Baseline)	6.32	-	0.305
The AVEC 2019 Dataset (Test)	Rodrigues *et al.*, 2019 [59]	7.02	-	0.199
	Zhang *et al.*, 2019 [60]	6.78	-	-
	Ringeval *et al.*, 2019 [62] (Baseline)	8.19	-	0.108

Table 3.3: Hyperparameter spaces for the learning algorithms

Algorithm	**Classification**		**Regression**	
	Feature Vector Per Utterance	**Feature-value Binning**	**Feature Vector Per Utterance**	**Feature-value Binning**
NN	batch_size: 16, 32, 128; epochs: 10, 50, 100; learn_rate: 0.001, 0.01, 0.1, 0.2, 0.3; momentum: 0.0, 0.2, 0.4, 0.6, 0.8, 0.9; dropout_rate: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5; neurons_per_layer: 16, 32,128, 256, 512, 1028	batch_size: 16,32,128; epochs: 10, 50, 100; learn_rate: 0.001, 0.01, 0.1, 0.2, 0.3; dropout_rate: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5; neurons_per_layer: 16, 32,128, 256, 512, 1028	-	-
HMM	n_components=2; n_iteration: 100, 300, 500	n_components=2; n_iteration: 50, 100, 300	-	-
DT	criterion: gini, entropy; max_depth: 3, 4, 6, 8, 10; max_leaf_nodes: 10, 20, 30; min_samples_leaf: 5, 7, 9, 11, 13, 15	criterion: gini, entropy; max_depth: 3, 4, 6, 8, 10; max_leaf_nodes: 10, 20, 30; min_samples_leaf: 5, 7, 9, 11, 13, 15	criterion: mse, friedman_mse, mae; max_depth: 3, 4, 6; max_leaf_nodes: 10, 20, 30; min_samples_leaf: 5, 7, 9	criterion: mse, friedman_mse, mae; max_depth: 3, 4, 6; max_leaf_nodes: 10, 20, 30; min_samples_leaf: 5, 7, 9
GBT	max_depth: 3, 5, 7, 9, 11, 13, 15: max_features: auto, sqrt; min_samples_leaf: 10, 20, 40; min_samples_split: 2, 5, 10, 20: n_estimators: 100, 300, 500; subsample: 0.5, 0.75, 0.9	max_depth: 3, 5, 7, 9, 11, 13, 15: max_features: auto, sqrt; min_samples_leaf: 10, 20, 40; min_samples_split: 2, 5, 10, 20: n_estimators: 100, 300, 500; subsample: 0.5, 0.75, 0.9	learning_rate: 0.01, 0.02, 0.05: max_depth: 3, 5, 7, 9; min_samples_leaf: 10, 20, 30; min_samples_split: 10, 20, 30; n_estimators: 100, 300, 500	learning_rate: 0.01, 0.02, 0.05: max_depth: 3, 5, 7, 9; min_samples_leaf: 10, 20, 30; min_samples_split: 10, 20, 30; n_estimators: 100, 300, 500
RF	max_depth: 5, 7, 9, 11, 13; min_samples_leaf: 10, 20, 40; min_samples_split: 2, 5, 10; n_estimators: 100, 300, 500, 800	max_depth: 3, 5, 7, 9, 11, 13, 15; min_samples_leaf: 10, 20, 40; min_samples_split: 2, 5, 10; n_estimators: 100, 300, 500	max_depth: 5, 10, 15; min_samples_leaf: 10, 20, 40; min_samples_split: 2, 5, 10; n_estimators: 100, 300, 500, 800	max_depth: 3, 5, 7; min_samples_leaf: 10, 20, 40; min_samples_split: 2, 5, 10; n_estimators: 50, 100, 300
SVM	C: 0.1, 1, 10, 100; gamma: 1, 0.1, 0.01, 0.001, scale; kernel: rbf, poly, sigmoid; max_iter: 300, 500, 800	C: 0.1, 1, 10, 100; gamma: 1, 0.1, 0.01, 0.001, scale; kernel: rbf, poly, sigmoid; max_iter: 50, 100, 300, 500	C: 0.1, 1, 10, 100; gamma: 1, 0.1, 0.01, 0.001, scale; kernel: rbf, poly, sigmoid; ’max_iter: 100, 300, 500, 800	C: 0.1, 1, 10, 100; gamma: 1, 0.1, 0.01, 0.001, scale; kernel: rbf, poly, sigmoid; ’max_iter: 50, 100, 300, 500
XGB	gamma: 0.001, 0.01, 0.1, 0.3; learning_rate: 0.005, 0.01, 0.05, 0.1; max_depth: 5, 7, 9; n_estimators: 300, 500, 800; subsample: 0.6, 0.8, 1.0	gamma: 0.001, 0.01, 0.1, 0.3; learning_rate: 0.005, 0.01, 0.05, 0.1; max_depth: 3, 5, 7; n_estimators: 50, 100, 300; subsample: 0.6, 0.8, 1.0	colsample_bytree: 0.6, 0.8, 1.0; gamma: 0.1, 0.3, 0.5; learning_rate: 0.005, 0.01, 0.05, 0.1; max_depth: 5, 10, 15; n_estimators: 100, 300, 500; subsample: 0.6, 0.8, 1.0	colsample_bytree: 0.6, 0.8, 1.0; gamma: 0.1, 0.3, 0.5; learning_rate: 0.005, 0.01, 0.05, 0.1; max_depth: 3, 5, 7; n_estimators: 50, 100, 300; subsample: 0.6, 0.8, 1.0

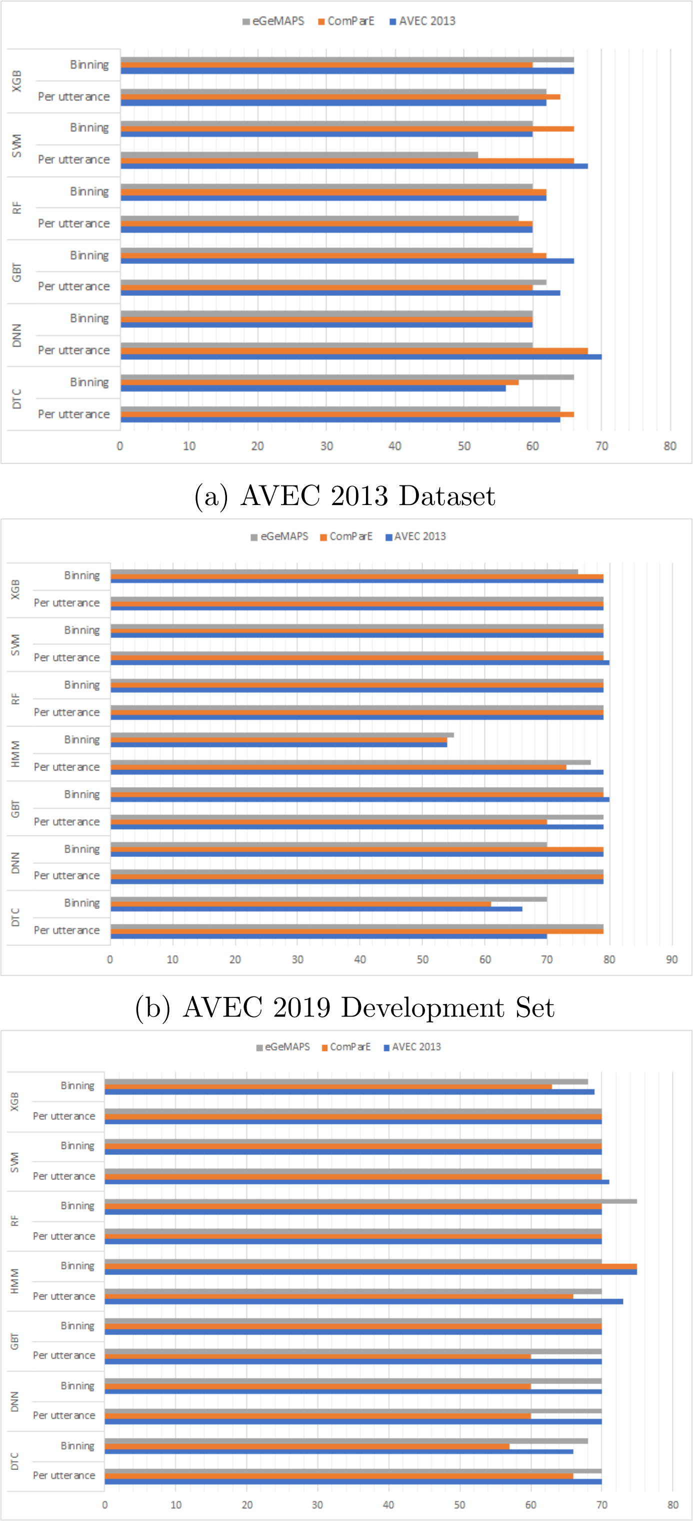
Table 3.4: Classification results. The best accuracy and F_1 scores for each data set are marked in bold. The worst accuracy of the same classifier has been underlined.

Feature set	**Algorithm**	**The AVEC 2013**		**The AVEC 2019 (Dev.)**		**The AVEC 2019 (Test)**	
		Per Utterance	**Binning**	**Per Utterance**	**Binning**	**Per Utterance**	**Binning**
		F_1(N.D)/F_1(D)/Acc. [*]	F_1(N.D.)/F_1(D)/Acc.	F_1(N.D)/F_1(D)/Acc.	F_1(N.D)/F_1(D)/Acc.	F_1(N.D)/F_1(D)/Acc.	F_1(N.D)/F_1(D)/Acc.
AVEC 2013	DTC	0.78/0.26/0.64	0.67/0.35/0.56	0.82/0.00/0.70	0.79/0.17/0.66	0.82/0.00/0.70	0.79/0.17/0.66
	DNN	**0.80**/0.40/**0.70**	0.74/0.09/0.60	0.88/0.00/0.79	0.88/0.00/0.79	0.82/0.00/0.70	0.82/0.00/0.70
	GBT	0.77/0.18/0.64	0.77/0.26/0.66	0.88/0.00/0.79	0.89/0.15/**0.80**	0.82/0.00/0.70	0.81/0.19/0.70
	HMM	0.70/0.41/0.60	0.78/0.26/0.66	0.87/0.33/0.79	0.67/0.24/0.54	0.84/0.29/0.73	0.83/**0.53**/**0.75**
	RF	0.75/0.00/0.60	0.75/0.17/0.62	0.88/0.00/0.79	0.88/0.00/0.79	0.82/0.00/0.70	0.82/0.00/0.70
	SVM	0.78/0.38/0.68	0.75/0.00/0.60	0.88/**0.48**/**0.80**	0.88/0.00/0.79	0.83/0.25/0.71	0.82/0.00/0.70
	XGB	0.76/0.10/0.62	0.76/0.41/0.66	0.88/0.00/0.79	0.88/0.14/0.79	0.82/0.00/0.70	0.80/0.18/0.69
ComParE	DTC	0.78/0.26/0.66	0.70/0.28/0.58	0.88/0.00/0.79	0.76/0.00/0.61	0.78/0.26/0.66	0.69/0.29/0.57
	DNN	0.78/0.38/0.68	0.75/0.00/0.60	0.88/0.00/0.79	0.88/0.00/0.79	0.75/0.00/0.60	0.75/0.00/0.60
	GBT	0.75/0.00/0.60	0.73/0.34/0.62	0.82/0.00/0.70	0.88/0.00/0.79	0.75/0.00/0.60	0.82/0.00/0.70
	HMM	0.75/0.48/0.66	-/-/-	0.83/0.35/0.73	0.67/0.24/0.54	0.77/0.34/0.66	0.83/**0.53**/**0.75**
	RF	0.75/0.00/0.60	0.75/0.17/0.62	0.88/0.00/0.79	0.88/0.00/0.79	0.82/0.00/0.70	0.82/0.00/0.70
	SVM	0.77/0.32/0.66	0.75/0.45/0.66	0.88/0.00/0.79	0.88/0.00/0.79	0.82/0.00/0.70	0.82/0.00/0.70
	XGB	0.76/0.25/0.64	0.72/0.29/0.60	0.88/0.00/0.79	0.88/0.14/0.79	0.82/0.00/0.70	0.76/0.09/0.63
eGeMAPS	DTC	0.75/0.36/0.64	0.75/0.48/0.66	0.88/0.00/0.79	0.81/0.26/0.70	0.82/0.00/0.70	0.77/0.47/0.68
	DNN	0.75/0.00/0.60	0.75/0.00/0.60	0.88/0.00/0.79	0.81/0.26/0.70	0.82/0.00/0.70	0.82/0.00/0.70
	GBT	0.75/0.24/0.62	0.75/0.00/0.60	0.88/0.00/0.79	0.88/0.00/0.79	0.82/0.00/0.70	0.82/0.00/0.70
	HMM	0.74/**0.51**/0.66	0.77/0.18/0.64	0.87/0.13/0.77	0.67/0.24/0.55	0.82/0.11/0.70	0.83/**0.53**/**0.75**
	RF	0.71/0.22/0.58	0.75/0.00/0.60	0.88/0.00/0.79	0.88/0.00/0.79	0.82/0.00/0.70	0.82/0.00/0.70
	SVM	0.75/0.24/0.52	0.75/0.00/0.60	0.88/0.00/0.79	0.88/0.00/0.79	0.82/0.00/0.70	0.82/0.00/0.70
	XGB	0.74/0.30/0.62	0.75/0.45/0.66	0.88/0.00/0.79	0.85/0.12/0.75	0.82/0.00/0.70	0.79/0.36/0.68

[*] F_1(N.D) = F_1 Score for non-depressed class; F_1(D) = F_1 Score for depressed class; Acc. = Accuracy

Classification Results: Table 3.4 shows the classification results applying the two preprocessing methods described in Section 3.3.3. The highest overall accuracy of 70% on the AVEC 2013 dataset has been achieved by NN applying Feature Vector Per Utterance preprocessing on the AVEC 2013 feature set al.though this overall accuracy is the same as the findings of Lopez-Otero et al.[15], the HMM trained on eGeMAPS features through the same preprocessing method achieves F_1 a score of 0.51 which is better than the NN (0.40) by 27.5%.

Similar performance was also observed by the HMMs on the AVEC 2019 dataset. On the development set, both the highest accuracy of 80% (male: 0.79, female: 0.81) and depressed class F_1 score of 0.48 was achieved by the pipeline consisting of AVEC 2013 features, Feature Vector Per Utterance preprocessing, and SVM classifier. However, the performance of the HMM trained on AVEC 2013 features (accuracy 79%, depressed class F_1 0.33) closely follows the SVM; and on the test set the HMM model shows better performance in terms of both accuracy (HMM: 73%, SVM: 71%) and depressed class F_1 scores (HMM: 0.29, SVM 0.25). The performance of the HMM on the test set shows further improvement when feature value binning is applied, raising the accuracy to 75% (male: 0.79, female: 0.62) and the F_1 on the depressed class to 0.53. The classification accuracies have been illustrated in Figure 3.2.



(a) AVEC 2013 Dataset

(b) AVEC 2019 Development Set

(c) AVEC 2019 Test Set

Figure 3.2: Classification accuracy

Table 3.5: Regression results. Best performing regressor has been marked in bold. Highest RMSE of the same regressor has been underlined

Feature set	Algorithm	The AVEC 2013				The AVEC 2019 (Dev.)				The AVEC 2019 (Test)			
		Per Utterance		Binning		Per Utterance		Binning		Per Utterance		Binning	
		CCC	RMSE	CCC	RMSE	CCC	RMSE	CCC	RMSE	CCC	RMSE	CCC	RMSE
AVEC 2013	DTC	0.064	11.84	0.098	14.6	-0.042	5.89	**0.333**	5.43	0.000	6.48	-0.155	8.18
	GBT	0.178	11.39	0.322	11.48	-0.032	5.80	0.125	5.53	0.008	6.41	0.131	6.43
	RF	0.126	11.66	0.319	10.80	-0.013	5.74	0.197	**5.37**	-0.002	6.44	0.044	6.56
	SVM	0.412	**10.29**	0.400	11.59	-0.074	5.98	-0.018	5.91	-0.003	6.48	0.020	7.04
	XGB	0.111	11.90	0.319	11.40	-0.077	5.96	0.099	5.80	0.012	6.51	-0.040	6.93
ComParE	DTC	0.099	11.73	0.202	13.10	-0.034	5.89	-0.070	6.69	0.007	6.42	**0.190**	7.20
	GBT	0.198	11.33	0.257	11.02	-0.025	5.82	0.071	5.83	0.001	6.43	0.110	6.24
	RF	0.075	11.84	0.257	11.02	-0.005	5.75	-0.029	6.02	0.000	6.43	0.074	**6.37***
	SVM	0.216	11.63	0.369	10.74	-0.046	5.96	-0.034	5.96	0.004	6.51	0.006	7.10
	XGB	0.146	11.67	0.287	11.49	-0.023	5.84	-0.083	6.40	0.032	6.38	0.072	6.83
eGeMAPS	DTC	0.135	12.00	0.341	13.32	-0.034	5.89	-0.070	6.69	0.000	6.46	0.031	6.98
	GBT	0.148	12.00	**0.431**	10.88	-0.025	5.82	0.071	5.83	0.011	6.40	0.044	6.55
	RF	0.109	11.67	0.244	11.26	-0.005	5.75	-0.029	6.02	0.021	6.37	0.049	6.49
	SVM	0.072	12.04	0.051	11.94	-0.046	5.96	-0.034	5.96	0.018	6.58	0.031	7.02
	XGB	0.113	12.23	0.410	11.14	-0.023	5.84	-0.083	6.40	0.042	6.43	-0.029	6.97

* Score better than state-of-the-art

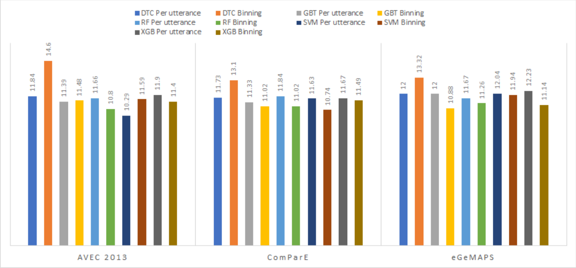


Figure 3.3: Regression performance (RMSE) on AVEC 2013 Dataset

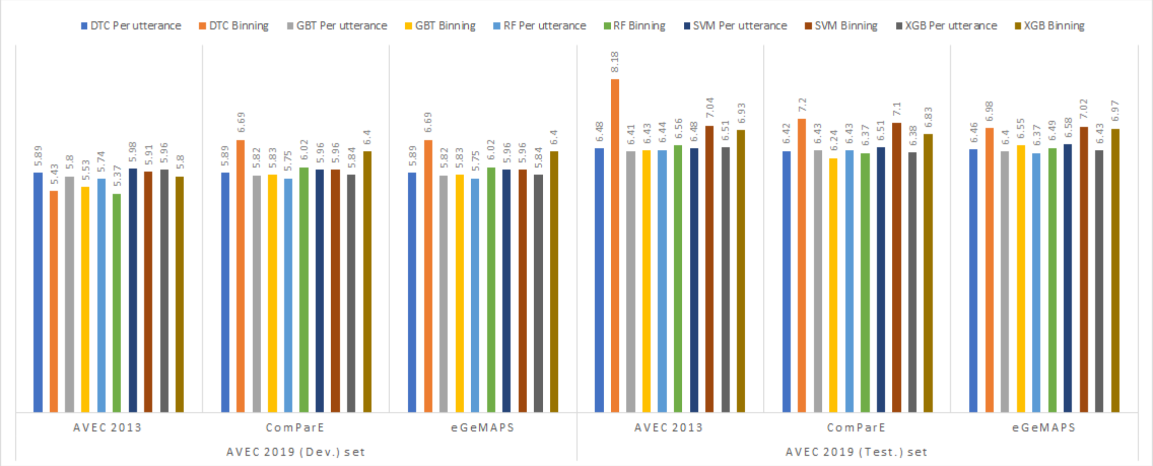


Figure 3.4: Regression performance (RMSE) on AVEC 2019 Dataset

Regression Results The SVM trained on AVEC 2013 features preprocessed using Feature Vector Per Utterance achieves RMSE of 10.29, which is lower than the baseline (10.75) on the AVEC 2013 dataset, although the Gaussian staircase methodology proposed by Williamson *et al.* achieves even lower RMSE of 8.68 on this dataset (3.5) [22].

Table 3.5 also shows that application of feature value binning achieves noticeable performance gain of the regressors in comparison to Feature Vector Per Utterance on the AVEC 2019 dataset. The lowest RMSE of 5.37 (CCC 0.197) was achieved by random forest on AVEC 2013 features, followed by decision tree (RMSE 5.43, CCC 0.197) on the development set. This score is a significant improvement from the baseline, and lower than Fan *et al.*, Zhang *et al.* and Rodrigues *et al.* (Table 3.2). Only the challenge winner Ray *et al.* reports a marginally lower RMSE of 5.11 on the development set. On the test set also, the lowest RMSE of 6.37 (CCC 0.074) was achieved by random forest applying feature value binning on ComParE features. But a higher CCC of 0.190 on the test partition was obtained by decision tree on the same preprocessed features, although compromising RMSE (7.20). The test RMSE of the random forest outperforms that of Ringeval *et al.* [62], Rodrigues *et al.* [59] and Zhang *et al.* [60].

Figure 3.3 and 3.4 depict the performance of the regressors on AVEC 2013 and AVEC 2019 datasets, respectively.

Performance comparisons against other techniques including the baseline has been presented in Table 3.6. As classification results were not reported on AVEC 2019 dataset in the literature and we could compare our classification accuracy only with Lopez-Otero *et al.*, the table shows comparison of regression error of our method against others.

Table 3.6: Performance comparison against others measured in RMSE, MAE and CCC

Dataset	**Method**	**RMSE**	**MAE**	**CCC**
The AVEC 2013 DSC Dataset	Valstar *et al.* [14] (challenge baseline)	10.75	8.66	-
	Williamson *et.al.*, 2013 [22]	8.68	7.12	-
	Jan *et al.*, 2017 [51]	10.08	8.25	-
	He and Cao, 2018 [23]	9.89	8.19	-
	Özkanca *et al.*, 2018 [24]	9.42	-	-
	Morales, 2018 [25]	10.70	8.59	-
	Zhao *et al.*, 2020 [52]	9.65	7.38	-
	Ours	10.29	8.3	0.400
The AVEC 2019 Dataset (Development)	Ringeval *et al.*, 2019 [62] (Baseline)	6.32	-	0.305
	Ray *et al.*, 2019 [58]	5.11	-	-
	Rodrigues *et al.*, 2019 [59]	5.70	-	0.497
	Zhang *et al.*, 2019 [60]	5.83	-	-
	Fan *et al.*, 2019 [61]	6.20		0.348
	Ours	5.37	4.34	0.197
The AVEC 2019 Dataset (Test)	Ringeval *et al.*, 2019 [62] (Baseline)	8.19	-	0.108
	Rodrigues *et al.*, 2019 [59]	7.02	-	0.199
	Zhang *et al.*, 2019 [60]	6.78	-	-
	Ours	6.37	4.95	0.074

Discussion

Let us now revisit the three research questions, driving our work.

How informative is audio as a modality to detect the prevalence (classification) and severity (regression) of depression?

According to the experimental results presented in Table 3.4, vocal acoustic biomarkers carry sufficient information to classify depressed versus non-depressed individuals with 70 to 80% respectively from two benchmark datasets in German and English language. However, the highest F_1 scores for the depressed class were 0.51 and 0.53 respectively for the two datasets, which implies that the model needs to be more sensitive and specific to be used in a real-world system. In the two datasets we have used, there is only one audio sample of each individual, and there is no overlap between training and test samples. Therefore, availability of time series and personalized data may further improve the model's accuracy in a practical system.

What is the appropriate feature set, effectively balancing output quality and computational cost?

For the classification task, AVEC 2013 feature set was found most effective on both datasets. For the regression task also, the lowest RMSE on the AVEC 2013 and the AVEC 2019 development set were achieved using this feature set. However, on the test partition of the AVEC 2019 dataset, both lowest RMSE (6.37) and highest CCC (0.190) were scored using ComParE dataset. But the improvement is marginal from AVEC 2013 feature set (RMSE 6.43, CCC 0.131) and is out-weighted by the additional computational cost of constructing the ComParE feature set (6373 features) in comparison to AVEC 2013 feature set (2268 features).

For both of the tasks, the application of feature value binning offers a significant reduction in computational complexity, with identical or better performance of the models. Particularly for the regression task, feature value binning improved the CCC score on both datasets. Moreover, the lowest RMSEs on both development and test partitions of the AVEC 2019 dataset were achieved using this method.

What is the most effective (type of) algorithm and configuration for the above two tasks?

Our experimental results show that HMMs are more effective than the other classifiers in diagnosing depression from this audio dataset. On both datasets, the HMMs achieved the highest F_1 scores for the depressed class on held-out sets, with marginal or no compromise on overall accuracy.

The lowest RMSE on the AVEC 2013 dataset was achieved by the SVM model which works better than other models on sparsely distributed small-sized datasets [88]. Random forest scored the lowest RMSE on both development and test partitions of the AVEC 2019 dataset. However, the highest CCCs were obtained by GBT and DTC respectively on the two datasets. Therefore, it is difficult to suggest the best regressor for the task. Dissimilarities between the two datasets, including language (German vs English), depression scale (BDI-II vs PHQ-8), nature and length of recordings (guided task vs interview), and bias towards lower scores are some possible reasons behind this phenomenon.

Recommendation: In order to recommend a data-processing-and-analysis pipeline for detecting depression in a real-world scenario, we need to develop a framework that exhibits consistent performance on both validation and test partition of the available data. In addition to overall accuracy, the model should have high sensitivity and specificity to the depressed class. Moreover, to derive such conclusions from the experiments described in this manuscript one should also consider that the two datasets used here are significantly different in terms of language, length and content, therefore one may consider to apply different methodology depending on the scenario.

Considering the above and the experimental results presented in this section, we can derive the conclusion that **HMM model trained on eGeMAPS features is the best for classification task**. Feature value binning provided best performance on comparatively longer English dataset (AVEC 2019), while on the shorter length German dataset application of Feature Vector Per Utterance performed better. In a real-world context, one would expect to collect longitudinal data and more varied/realistic than reading a short book passage; therefore we are inclined to recommend **"HMM model trained on eGeMAPS features with feature value binning"** for classification.

Keeping the same facts in mind, **for the regression task AVEC 2013 featureset**

preprocessed through Feature Binning should be chosen to train **random forest regressor**.

By no means, does our experiment conclusively recommend these methodologies; more research and comparative analysis is necessary on benchmark and real-world data. Nevertheless our experiments provide a strong signal in this direction.

3.6 Conclusion

In this paper, we have systematically investigated the performance of different classifiers and regressors, for detecting depression from speech by analyzing acoustic features. Our experiments, with the AVEC 2013 and 2019 German and English datasets, demonstrate that our feature-value binning preprocessing method leads to a noteworthy improvement in the performance of the regressors and competitive classification performance, while also significantly reducing the computational complexity of model training. Although HMM classifiers were found most effective in distinguishing "depressed" samples, the choice of feature set was subject to the characteristics of the datasets. Several regressors performed competitively in predicting depression severity. In the future, we plan to further investigate feature selection, especially studying deep-learned features from audio spectrograms and the interplay between features and language.

Chapter 4

Cost-effective Models for Detecting Depression from Speech

In this experiment, I analyzed the newly curated depression and anxiety corpus, DEPAC [89] to determine how the proposed depression prediction pipeline performs in real-world scenarios.

- **Data preprocessing:** Unlike the AVEC datasets, the samples of the DEPAC dataset were not captured in a controlled environment. Instead, they were recorded on the participants' personal devices. Therefore, background noise was expected in these speech samples. To deal with that, I implemented an audio enhancement step (Figure 1.2) of the pipeline. As features, I considered a popular deep representation feature set VGG-16, and compared its effectiveness with a hand-curated conventional acoustic feature set.

- **Model training:** Three machine learning algorithms, i.e., SVM, random forest, and feedforward neural network (FNN), were trained in this experiment. The models were selected based on previous experience and existing literature.

- **Predictions:** I reported the findings on the effect of relevant variables such as content and length of the samples, and depression severity of the subject on the models' predictions. I also identified the pros and cons of each kind of feature in terms of prediction error and computational cost using 5 subject independent cross-validation folds.

Key findings: The key findings of this experiment can be listed as follows:

1. The **audio enhancement** step significantly affected the values of 94% of the 220 conventional acoustic features used in this experiment.

2. SVM and FNN models performed better on **manually curated conventional feature** than on VGG-16, at a remarkably less computation time. The performance of random forest was marginally better (0.0004%) on VGG-16. Therefore, the application of **deep-representation features** in predicting depression severity would require further investigation and refinement.

3. **Content** and **length** of speech samples did not show a significant impact as long as the length of speech samples is reasonably short, less than one minute in the case of DEPAC dataset.

4. **Sex** and **depression severity** of the subject were found to correlate with the prediction accuracy.

This chapter has been compiled based on our paper published in the 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1687-1694). IEEE.

As the first author, I conducted the experiments under the supervision of Jekaterina Novikova during my internship at the Winterlight Labs (acquired by Cambridge Cognition in 2023)

Cost-effective Models for Detecting Depression from Speech

Abstract: Depression is the most common psychological disorder and is considered a leading cause of disability and suicide worldwide. An automated system capable of detecting signs of depression in human speech can contribute to ensuring timely and effective mental health care for individuals suffering from the disorder. Developing such an automated system requires accurate machine-learning models, capable of capturing signs of depression. However, state-of-the-art models based on deep acoustic representations require abundant data, meticulous selection of features, and rigorous training; the procedure involves enormous computational resources. In this work, we explore the effectiveness of two different acoustic feature groups — conventional hand-curated and deep representation features, for predicting the severity of depression from speech. We explore the relevance of possible contributing factors to the models' performance, including the gender of the individual, the severity of the disorder, content, and length of speech. Our findings suggest that models trained on conventional acoustic features perform as well as, or better than the ones trained on deep representation features at significantly lower computational cost, irrespective of other factors, e.g., content and length of speech, gender of the speaker, and severity of the disorder. This makes such models a better fit for deployment where the availability of computational resources is restricted, such as real-time depression monitoring applications in smart devices.

4.1 Introduction

Depression is a common psychological disorder. About 264 million people worldwide suffer from depression, which is almost 5% of the world's total population [12]. Only about 50% of the people experiencing major depression receive treatment. Due to the lack of continuous monitoring and timely support, depression causes one death every 40 seconds, resulting in 800,000 deaths by suicide worldwide every year [12].

Conventional depression diagnostic systems, such as clinical assessment or standard questionnaires, require a significant amount of time and active participation of the depressed individuals. Studies reveal that depression is reflected in behavioural fluctuations of certain day-to-day activities and physical parameters [13]. These findings have accelerated interventions in depression recognition using predictive models that incorporate input data of different modalities, among which audiovisual is one of the most explored areas. In this work, we emphasize audio modality for its manifold benefits. Audio-based depression detection systems offer better privacy for users of remote monitoring systems. This kind of automated assessment takes only a few minutes of audio recording, therefore is less time-consuming, and would reduce the burden on the individuals.

Multiple research efforts aim to develop a system that detects depression by analyzing the fluctuation of acoustic features in human speech [90], [71]. An ML model that detects evidence of depression from audio data is a prerequisite for such a system. Existing best-performing ML models that detect mental and cognitive diseases from audio data use either deep representation acoustic features, or a combination of conventional hand-crafted and deep features [58, 91]. Although deep representation features offer a unified process of feature extraction, feature selection, and model training, extracting and processing these features demands enormous computation resources including memory and processing time. This makes such models inconvenient for many real-world applications, where speed of data processing, model training, and inference are of crucial importance [92]. Therefore, researchers and system designers need to make a choice of features when developing and deploying the model, considering both performance and cost. Some previous research compares the two approaches in the domain of cognitive disease detection [93] but to the best of our knowledge, no such research has been done so far in the domain of depression. To address this gap, in this work, we have experimented with both hand-crafted conventional acoustic features and deep representation acoustic features. We address the following research questions:

1. *Between conventional and deep representation acoustic features, which ones are more effective in determining depression severity in terms of accuracy and computational cost?*
2. *Does the machine learning (ML) model performance vary based on the gender of the subject?*
3. *What is the effect of content and length of speech data in predicting depression from speech?*

Answers to these questions enable the research community as well as system designers to make informed choices of modality, features, and algorithms that suit best to the context, e.g. target user group and affordability. In this work, we compare the performance of the ML models trained on each type of feature extracted from speech samples of a variety of content and length. Our key findings suggest that:

1. ML model trained on conventional acoustic feature set curated using expert domain knowledge demonstrates competitive performance as state-of-the-art models in predicting depression severity, irrespective of length and content of speech, and gender of the speaker
2. Usage of deep representation features resulted in marginal improvement of performance (0.0004%) consuming 1000 times more memory and 3000 times more computation time.

As such, we claim that models predicting depression from human speech that are trained on conventional acoustic features are a better choice than the models trained on deep acoustic representations in situations when computational resources are limited, e.g. in mental healthcare applications for portable or wearable devices. On the other hand, deep representation models fit better to the scenarios where abundant training data is available, for example, social media, and computational resources are a legitimate trade-off for better performance.

4.2 Related Works

Individuals suffering from psychological and neurological disorders like depression exhibit measurable fluctuation in vocal parameters ([35] and [36]). A significant number of research have been conducted to relate these parameters with depression severity. DAIC-WoZ dataset [40] is a widely used dataset in acoustic-based depression severity prediction, consisting of structured interviews of participants conducted by a virtual agent. Two subsets of this dataset have been introduced as the challenge corpus of three Audio/Visual Emotion Challenges (AVEC) in 2016 [94], 2017 [16] and 2019 [62], where participants proposed machine learning models to predict depression score on the PHQ-8 scale [95]. Handcrafted acoustic features have been exploited for this task for the last few decades, while deep representations of acoustic features have become popular in recent years. Further, we present a summary of existing works in this area and compare them based on the type of acoustic features.

4.2.1 Conventional Acoustic Features

Conventional acoustic features fall in temporal, spectral, energy, and voicing-related categories, from which researchers hand-pick the ones that are most suitable for predicting certain disorders, such as depression [36]. Over time, certain sets of these acoustic features, introduced in speech emotion and depression recognition challenges, have gained popularity, among which baseline feature sets of AVEC 2013 [14] and AVEC 2016 [94], INTERSPEECH ComParE [72], extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [73] are noteworthy. Development of feature extraction toolkits like openSMILE [85], and COVAREP [65] has made it easier for researchers to extract these features for speech analysis in different aspects.

4.2.2 Deep Representation Features

Deep representations of acoustic features are inspired by the deep learning paradigms common in image processing. Here, spectral images of speech instances are fed into pre-trained image recognition CNNs, and a set of the resulting activations are extracted as feature vectors. In AVEC 2019 Depression Detection Sub-challenge (DSC), deep representation features from four robust pre-trained CNNs using VGG-16 [46], AlexNet [96], DenseNet-121, and DenseNet-201 [67] were included as challenge baseline features. Participants chose between using one or more sets of deep representation features ([59], [66]), and combining them with traditional features [58] and obtained competitive performances (Table 3.6). Deep representation provides the option to unite feature extraction, feature selection and model training into a single automated generalisable procedure, compromising the opportunity to incorporate expert domain knowledge [97], and necessitating considerably higher computational cost.

4.2.3 Depression Detection Models

AVEC 2016 challenge dataset was used in analysis presented in [53], [26], [30], [28], [29], [54], [55], [98] and [49]. Williamson *et al.* extracted formants, MFCCs, glottal features, loudness [53]. In addition to COVAREP [65] audio features, [26] took text topics into account, while [30] considered a more extended set of features of audio, video and text modalities. [29] added Delta and Delta-Delta coefficients, mean, median, standard deviation, peak-magnitude to RMS ratio to the set of challenge baseline audio features. Applying similar higher-order statistics on the baseline features, [54] constructed an extended feature set of 553 features, of which they identified 279 features with statistically significant univariate correlation. [55] implemented multi-modal sentence-level embedding on log-Mel spectrogram and MFCC features. In their recent work, [56] exploited a combination of eGeMAPS and INTERSPEECH features extracted from the longest ten segments of each audio sample. They reshape the feature vector in an image-like 2D feature map in row-major order and adopted Deep Convolutional Generative Adversarial Net (DCGAN) for feature vector generation. [49] trained three spectrogram-based Deep Neural Network architectures phoneme consonant and vowel units and their fusion. Their findings suggest that deep learned consonant-based acoustic characteristics lead to better recognition results than vowel-based ones, and the fusion of vowel and consonant speech characteristics outperforms the other models on the task. [98] described a transfer attention mechanisms from speech recognition to aid depression severity measurement. The transfer is applied in a two-level hierarchical network which reflecting the natural hierarchical structure of speech.

The AVEC 2016 challenge corpus included training, development and test partitions of audio samples. Using acoustic features exclusively, the lowest root-mean-square-error (RMSE) of 5.52 and 6.42 on the development and test set were reported in [99] and [57] respectively. RMSE 4.99, 5.40, 5.66 and 6.42 were reported in [30], [99], [98] and [57] respectively on the test set. The challenge baseline RMSE 6.74 (mean absolute error (MAE) 5.36) and 7.78 (MAE 5.72) were set for the development and test set respectively [16].

Comparatively fewer studies have been conducted on the AVEC 2019 DDS dataset; which is a super-set of the AVEC 2016 dataset. A wide range of audio features has been provided as the challenge baseline encompassing both handcrafted sets of conventional features and deep representation of acoustic features, including eGeMAPS, Mel Frequency Cepstral Coefficients (MFCCs), Bag of Audio Words (BoAW) and two sets of deep spectrum features by feeding spectral images of speech instances into pre-trained image recognition Convolutional Neural Networks (CNN) (VGG-16 [67] and DenseNet-121 [46]) and extracting the resulting activations as feature vectors. Of these, deep spectrum features were used by [66] and [59], and MFCCs and eGeMAPS by [61]. [58] exploited all the baseline feature sets, while [60] extracted the AVEC 2017 baseline feature set [16] using the COVAREP software toolbox [65], in addition to eGeMAPS. Different configurations of long short-term memory (LSTM) networks were used in all of these works except Zhang *et al.* [60], who adopted random forest and logistic regression. Acoustic models proposed by [58] and [60] achieved lowest RMSE of 5.11 and 6.78 on the development and test partitions, respectively.

4.3 Materials and Methods

4.3.1 Dataset

We use DEPression and Anxiety Crowdsourced corpus (DEPAC) [89] in this experiment, which consists of 2,674 audio samples collected from 571 subjects located in Canada and the United States. 54.67% of the study subjects are female and 45.33% are male, aged between 18 and 76 years, and their formal education ranged from 1 to 26 years. The data was collected via crowdsourcing and consists of a variety of self-administered speech tasks (Table 4.1). The participants completed these tasks using Amazon Mechanical Turk (mTurk) [1]. The speech tasks were curated to increase the phonemic variety and were supported by literature on detecting mental disorders, such as Alzheimer's Disease (AD) [8] and depression [9], [10], [11] from speech.

Table 4.1: Speech tasks in DEPAC corpus

Speech Task	**Description**	**Average Duration**
Phoneme Task	Record "aah" sound for as long as the participant could hold breath	5.79 sec
Phonemic Fluency	Pronounce as many unique words as possible starting with the letters "F", "A" or "S"	22.13 sec
Picture Description	Describe a picture shown on the screen (example Figure 4.1)	46.60 sec
Semantic Fluency	Describe a positive experience they expected to have within five years in the future	43.76 sec
Prompted Narrative	Tell a personal story, describing the day, a hobby, or a travel experience	45.34 sec

In this dataset, the depression severity is represented by Patient Health Questionnaire (PHQ-9) scores, which is a 3-point self-rated measure for depressive symptoms, including 9 questions. To ensure comparability of our results with works done on popular subsets of DAIC-WoZ corpus [40] i.e., AVEC 2017 [16] and AVEC 2019 [62] we used responses to 8 PHQ questions in our analysis and reported our results on PHQ-8 scores. The score ranges from 0 to 24 on the PHQ-8 scale where a score in

[1] https://www.mturk.com

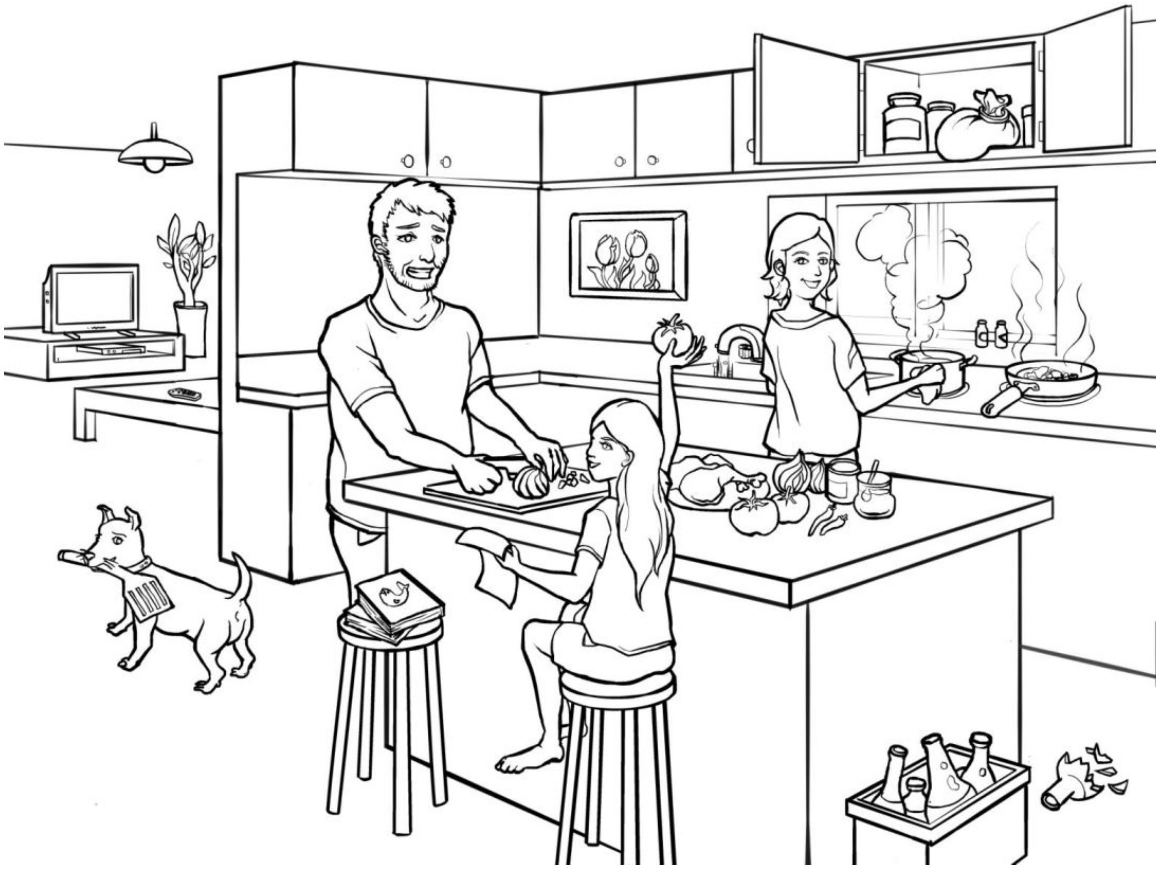


Figure 4.1: 'Family in the kitchen' image used in the picture description task.

the range (6,9), (10, 14), (15, 28) represents mild, moderate, and severe levels of depression respectively. The mean PHQ-8 score of DEPAC corpus (M) is 6.56 with a standard deviation (SD) of 5.56.

4.3.2 Audio Quality Enhancement

To suppress possible background noise present in the samples and improve the quality of the audio, we applied *logmmse* enhancement technique [100] on the audio samples. This method was found the best among existing audio enhancement algorithms in literature [101]. The enhancement step is found statistically significant ($p \leq 0.005$) on 94% of the 220 conventional acoustic features in the Wilcoxon signed-rank test with Bonferroni correction.

Audio volume was normalized to -20 dBFS across all speech segments to control for variation caused by recording conditions such as microphone placement.

4.3.3 Acoustic Features

We extracted two sets of acoustic features, representing hand-crafted sets of conventional features and deep learning features:

Conventional acoustic features

This set included 220 acoustic features, extracted from each audio sample. The feature set includes:

- **Spectral features:** Intensity (auditory model based), MFCC 0-12, Zero-Crossing Rate (ZCR)
- **Voicing-related features:** Fundamental frequency (F_0), Harmonic-to-Noise Ratio (HNR), shimmer and jitter, durational features, pauses and fillers, phonation rate

Statistical functionals including minimum, maximum, average, and variance were computed on the low-level descriptors. Additionally, skewness and kurtosis were calculated on MFCCs, first and second-order derivatives of MFCCs, and Zero Crossing Rate (ZCR) [102].

A Python implementation of the Praat phonetic analysis toolkit [103] has been used to extract the majority of these features. The MFCC features and their functionals were computed using `python_speech_features`[2] library.

Deep Representation Features

Deep representations of acoustic features are inspired by the deep learning paradigms common in image processing. Here, spectral images of speech instances are fed into pre-trained image recognition CNNs, and a set of the resulting activations are extracted as feature vectors. VGG-16 is a type of Convolutional Neural Network (CNN). We used the DeepSpectrum library [104] to extract features from a pre-trained VGG-16 CNN [67]. The speech files are first transformed into mel-spectrogram images with 128 mel-frequency bands. Then, the spectral images are forwarded through the pre-trained networks. A 4,096-dimensional feature vector is then formed from the activations of the second fully connected layer in VGG-16. The features were extracted at a window width of 1s and a hop size of 300 ms from each audio sample.

4.3.4 Data Preprocessing

Standardization

The range of values of audio features tends to vary widely. To ensure the even contribution of all features in the regression task, and to speed up gradient descent convergence of the deep neural network, once acoustic features were extracted from the audio samples, we standardized them using z-scores, i.e., subtracting the mean and dividing by standard deviation. The standard score of a sample x of feature f_i is calculated as:

$$z = \frac{x - \mu}{\sigma} \tag{4.1}$$

here μ and σ are the mean and standard deviation of the values of f_i in all training samples.

[2] https://pypi.org/project/python_speech_features/

Feature Selection

We applied the minimum Redundancy-Maximum Relevance (mRMR) algorithm [105] to select the most relevant features to the PHQ scores, minimizing redundancy in the selected set of features. 10% features were selected from the training set of each fold to train the ML models.

4.3.5 Model Training

Following [93] and [71], we train an array of linear and non-linear ML models separately on conventional and deep learning acoustic features:

- Support Vector Machines (SVM): Radial Basis Function (RBF) kernel SVM was trained. Values of hyperparameters *'C'* and *'gamma'* were tuned by 5-fold grid-search cross-validation (cv).
- Random Forest (RF): Scikit Learn implementation of random forest regressor was used. The number of estimator trees and maximum depth were tuned through grid-search cv.
- Feedforward Neural Network (FNN): The FNN model consists of 4 hidden layers, with 500 hidden units on the first layer, 250 in the second, and 125 in the rest of the hidden layers. 30% dropout on the output of each of the hidden layers of the FNN. We use the Adam optimizer in all FNN models with a learning rate of 0.001. Each of the FNN models is trained for 150 epochs.

The discussion presented by Balagopalan *et al.* [93] suggests that for small audio corpus like the ADReSSo challenge dataset (237 samples) [106], either leave-one-subject-out cross-validation or k-fold cross-validation can be applied. However, the dataset used in this work is considerably larger than the ADReSSo challenge dataset. Considering the size of the dataset and corresponding computational complexity, we decided to report evaluation metrics with 5-fold cross-validation (CV) for the models. We create 5 subject-independent folds, train the model using 4 of them, and use the rest for testing. The hyperparameters of the ML models were tuned using a 5-fold CV within each training set. We repeat the process for all 5 folds and report evaluation metrics averaging across predictions on all the folds. These folds preserve the same ratio of depression severity in each training and test partition.

To understand the effect of speech content on ML models' performance, we separated samples with each type of speech task and trained models on each type of them. We repeated the same process for conventional and VGG-16 features.

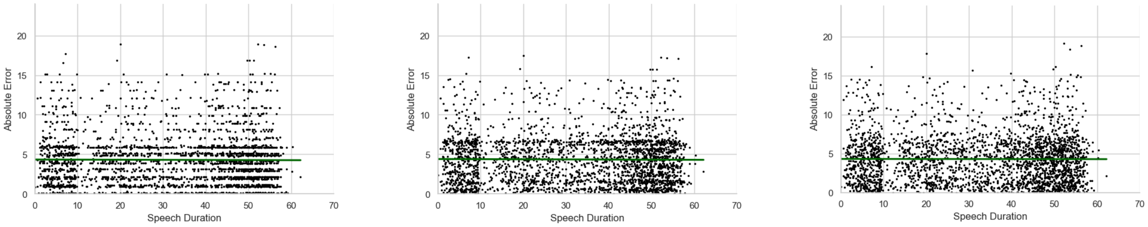
We ran our experiment on a MacBook Pro with an Intel Core i7 CPU at the clock speed of 2.67 GHz. The system availed 16 GB memory. The data preprocessing and model training was done in the Python programming language.

4.4 Result and Discussion

Here we present the performance of the ML models trained on the DEPAC dataset with a view to finding answers to the research questions (RQs), outlined in Section

Table 4.2: Regression error of models trained on conventional and VGG-16 features

Algorithm	**Sex**	**RMSE**		**MAE**	
		Conventional	VGG-16	Conventional	VGG-16
SVM	Male	5.04	7.89	4.22	6.95
	Female	5.64	7.11	4.33	6.23
	Overall	5.38	7.48	4.28	6.56
RF	Male	5.15	5.06	4.37	4.27
	Female	5.47	5.51	4.34	4.32
	Overall	5.32	5.31	4.31	4.33
FNN	Male	5.10	5.19	4.45	4.30
	Female	5.54	5.67	4.51	4.34
	Overall	5.35	5.46	4.40	4.32



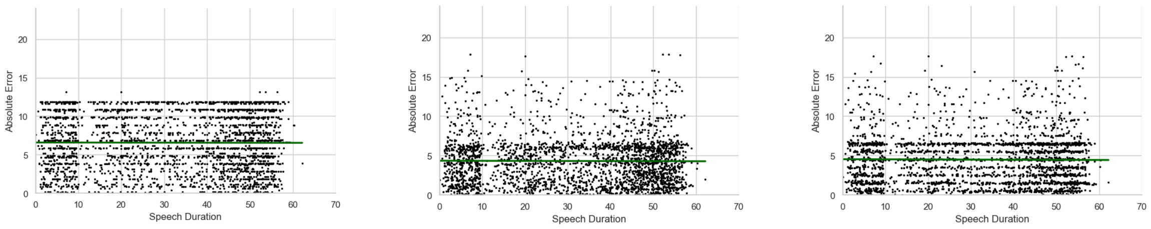
(a) SVM (CCC = 0.000) (b) RF (CCC = 0.003) (c) FNN (CCC = -0.002)

Figure 4.2: Correlation between speech length and prediction error of models trained on conventional acoustic features

4.1.

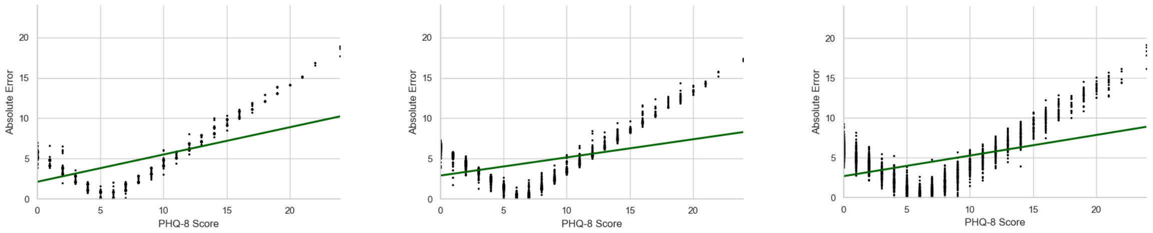
4.4.1 Effectiveness of different types of acoustic features in measuring depression

We trained 3 different ML models on each type of acoustic feature, i.e. conventional and VGG-16. We report the RMSE and MAE error of each model trained separately on samples from male and female subjects, along with the overall performance on the entire dataset (Table 4.2). We compare the performance of our best model with the state-of-the-art (Table 3.6). We report the CPU time required to train each model to assist future researchers and system designers in making informed choices of feature type and ML model.



(a) SVM (CCC = -0.004) (b) RF (CCC = -0.006) (c) FNN (CCC = -0.010)

Figure 4.3: Correlation between speech length and prediction error of models trained on VGG-16 features



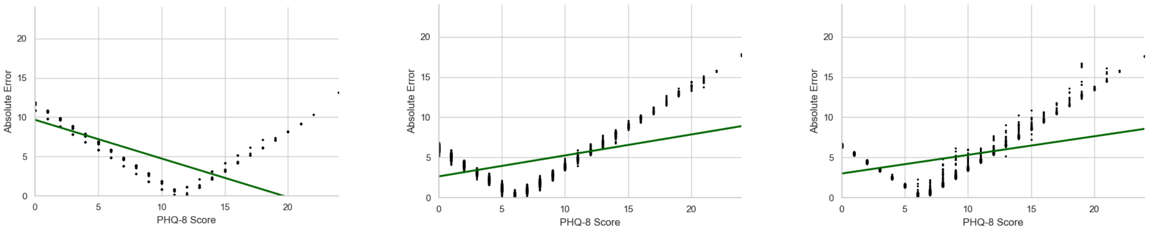
(a) SVM (CCC = 0.055) (b) RF (CCC = 0.385) (c) FNN (CCC = 0.435)

Figure 4.4: Correlation between depression severity and prediction error of models trained on conventional acoustic features

Performance of models trained on conventional and deep representation features

SVM and FNN models performed better on conventional features than on VGG-16, while the performance of RF is marginally better (0.0004%) on VGG-16 (Table 4.2). These findings are consistent with the previous works presented in Table 4.3. Conventional features presented in [56] modelled depression marginally better than models with deep representation features [59], [98].

In comparison to the state-of-the-art acoustic models, our proposed RF models show competitive performance. The RF model trained on both types of features outperforms almost all the existing works reporting similar performance metrics on the PHQ-8 scale (see Table 4.3). Only Ray *et al.* reported lower RMSE than us, fusing all four sets of AVEC 2019 baseline features, which is a combination of conventional (MFCC, Bag-of-Audio Words, eGeMAPS) and deep representation (VGG-16) features, and formulating a multi-level LSTM architecture [58]. Our proposed RF model trained on conventional features produces competitive performance to their proposed model, while substantially decreasing computational requirements. The VGG-16 features collected in the same manner as described by [58] from our audio corpus occupy 11.21 GB of memory, while our presented conventional feature file size is only 11 MB. Preprocessing and training models on conventional acoustic features took on average 3 minutes, while the procedure on VGG-16 features took at least 150 hours on the same computational environment (2.6 GHz 6 core Intel Core i7 processor, 16 GB memory). In short, our RF model using VGG-16 features offers 0.0004% improvement in performance



(a) SVM (CCC = -0.726)

(b) RF (CCC = 0.445)

(c) FNN (CCC = 0.394)

Figure 4.5: Correlation between depression severity and prediction error of models trained on VGG-16 features

than the same model using conventional features, using 1000 times more memory and 3000 times more processing time, implying similar or more computational resources is required for training complex models on multimodal features for marginal performance improvement. Therefore, the conventional features provided a better opportunity to adjust model parameters for performance improvement.

Results (Table 4.3) demonstrate that, in most cases, RMSE and MAE are lower for male subjects than female subjects. The reason behind this can be the lower severity of depression among male subjects than females in the DEPAC dataset [89]. The skewness in the dataset causes bias in model prediction, as described in [107]. For real-world applications, this issue needs to be taken care of by ensuring gender balance in training data.

Significance of performance deviation of models trained on conventional and deep representation features

We performed two-sample t-tests to identify if the performance deviations of the models are significant when trained on conventional acoustic features and VGG-16 features.

There was a significant difference between absolute errors in predictions of SVM models trained on WLL acoustic features ($M = 4.28, SD = 3.25$) and VGG-16 features ($M = 6.5, SD = 3.59); t(5332) = -24.20, p = 6.86e - 123 < .05$. The absolute errors in predictions of SVM model are significantly higher when trained on VGG-16 features than when trained on conventional features.

On the other hand, there was no significant difference between absolute errors in predictions of our best-performing RF and FNN models trained on conventional acoustic features ($RF : M = 4.34, SD = 3.09; FNN : M = 4.32, SD = 3.14$) and VGG-16 features ($RF : M = 4.31, SD = 3.11; FNN : M = 4.48, SD = 3.11$). The test scores of RF ($t(5332) = 0.38, p = .70 > .05$) and FNN ($t(5332) = -1.81, p = .07 > .05$) indicate that the deviation of errors in the prediction of the models are not significantly different irrespective of training features.

4.4.2 Effect of speech task type on ML model performance

The results (Table 4.5) do not reflect any significant deviation of model performance on the basis of speech task, therefore it is possible to recommend as a design choice any speech task of a similar length and content.

4.4.3 Correlation between model performance and speech length

From Figure 4.2 and 4.3 one can see that no significant correlation is observed between model performance (absolute error of each prediction) and the length of the corresponding sample. The Concordance Correlation Coefficient (CCC) scores for SVM, RF, and FNN models are 0.000, 0.003, and -0.002 for conventional features and -0.004, -0.006, and -0.010 for VGG-16 features respectively. The near-zero CCC values indicate that in the case of our dataset, the speech length of samples does not influence the models' performance. Note that all speech samples in DEPAC are less than one minute.

4.4.4 Correlation between depression severity and model performance

Absolute errors for each sample are plotted against ground truth PHQ-8 score for the models trained on conventional and VGG-16 features in Figure 4.4 and 4.5. The CCC scores for SVM, RF, and FNN models are 0.055, 0.385, and 0.435 for conventional features and -0.726, 0.445, and 0.394 for VGG-16 features respectively. The plots, along with high positive CCC scores for most of the models, imply that the samples with higher PHQ-8 scores contribute more to the overall prediction error of the models. This is caused by the imbalance in the number of samples with high and low PHQ-8 scores in the DEPAC dataset [89]. The higher density of samples with subthreshold (≤ 5) PHQ-8 score biases the models to make predictions close to the mean PHQ-8 (6.56) of the dataset. This observation strengthens the necessity of balancing the samples in training models to be used in real-world applications.

4.5 Conclusion and Future Works

Speech has proven to be a reliable marker for depression assessment. But in order to deploy a machine learning model in a practical system, it is necessary to identify the most informative acoustic feature, along with an efficient and cost-effective process to train the model. In this paper, we study the performance of conventional acoustic feature-based and pre-trained deep representation-based models on predicting depression severity from speech. We observe that the hand-curated feature-based approach achieves better performance in terms of lower RMSE and MAE, at a remarkably less computation time. Our experiments show that the gender of the speaker and distribution of score affect the model performance, and should be taken care of while formulating balanced training data. We also report that content and length of speech do not show a significant impact as long as the length of speech samples is reasonably short, less than one minute in our case. To summarize, we suggest using ML models trained on conventional features in resource-limited real-time situations and deep models in scenarios where fine-grained analysis involving higher computational power is crucial. In our future work, we plan to explore the generalizability of the findings across other datasets and disorders.

Table 4.3: Comparison of performance of SOTA ML models trained on different combinations of features. Bold denotes regression error of our proposed model.

Feature Type	**Study**	**RMSE**	**MAE**
Conventional	Formants, MFCCs, glottal features, loudness, AVEC 2017 dataset [22]	6.38	5.32
	COVAREP feature set, AVEC 2017 dataset [57]	6.34	5.30
	COVAREP features and functional, AVEC 2016 dataset)	6.50	5.13 [54]
	MFCC, AVEC 2016 dataset [55]	5.78	-
	MFCC and eGeMAPS features, AVEC 2019 dataset [61]	6.20	-
	eGeMAPS, INTERSPEECH features, AVEC 2016 dataset [56]	5.52	4.63
	eGeMAPS and COVAREP features, AVEC 2019 dataset [60]	6.78	5.77
Deep representation	VGG-16 features, AVEC 2019 dataset [59]	5.70	-
	Mel-spectra, AVEC 2017 dataset [98]	5.66	4.28
Conventional + deep combined	MFCC, BoAW, eGeMAPS and VGG-16 features, AVEC 2019 dataset [58]	5.11	-
Conventional	MFCCs, HNR, jitter, shimmer, ZCR features, DEPAC dataset [89]	**5.31**	**4.33**

Table 4.4: Time elapsed in different stages of model training

Processing step	**Algorithm**	**Conventional**	**VGG-16**
Data loading	-	1.132	244450
Preprocessing	-	96.834	545483
Model training	SVM	1.600	5593
	RF	0.715	981
	FNN	220.853	10270
Prediction	SVM	0.412	53
	RF	0.040	9
	FNN	1.102	7
Total	SVM	99.978	795579
	RF	98.715	790923
	FNN	267.227	800210

Table 4.5: Regression error of models trained on speech samples of different tasks

Algorithm	**Speech task**	**RMSE**		**MAE**	
		Conventional	VGG-16	Conventional	VGG-16
SVM	Semantic fluency	5.30	6.62	4.21	5.77
	Prompted narrative	5.29	6.63	4.24	5.77
	Phoneme task	5.49	6.40	4.33	5.54
	Phonemic fluency	5.45	6.49	4.35	5.62
	Picture description	5.43	6.54	4.36	5.67
RF	Semantic fluency	5.24	5.24	4.25	4.24
	Prompted narrative	5.31	5.25	4.30	4.25
	Phoneme task	5.39	5.29	4.37	4.30
	Phonemic fluency	5.38	5.29	4.38	4.30
	Picture description	5.42	5.31	4.40	4.31
FNN	Semantic fluency	5.34	7.13	4.28	5.34
	Prompted narrative	5.30	7.13	4.30	5.35
	Phoneme task	5.49	7.33	4.39	5.50
	Phonemic fluency	5.39	7.30	4.33	5.47
	Picture description	5.45	7.29	4.45	5.47

Chapter 5

A Machine-Learning Model for Detecting Depression, Anxiety, and Stress from Speech

The DEPAC dataset introduced in Chapter 4 contains multiple samples for each subject. However, these samples were recorded in a single session, as a result, this dataset did not reflect sufficient longitudinal variety. As a next step, the YouthDASS dataset was collected in collaboration with the research group at Tec de Monterrey, Mexico. A preliminary experimentation on the YouthDASS dataset is presented in this chapter.

- **Data preprocessing:** In this analysis, audio enhancement was applied, consisting of suppressing background noise and normalizing volume. The samples were divided into 5 *subject independent* folds maintaining a consistent ratio of disorder severity, holding out 20% samples for testing in each fold. 4,096 dimensional VGG-19 features from spectrograms of the audio samples were extracted using a pre-trained CNN model.
- **Model training:** A one-dimensional (1D) CNN model was trained on the VGG-19 features to predict the severity of each disorder on DASS-21 scale.
- **Predictions:** The trained 1-D CNN made predictions on the test partition of each of the 5 folds. Performance metrics were reported by averaging the predictions of the folds.

Key findings:

1. This work introduces YouthDASS, a new longitudinal multilingual speech corpus for depression, anxiety, and stress. The dataset captured valuable information on the post-pandemic effect on the mental health of youths.
2. The analysis of YouthDASS dataset validated that individuals with lower levels of depression, anxiety, and stress exhibit more conformity with routine activities,

demonstrated by a positive correlation between DASS-21 scores and adherence to our data collection protocol.

3. Finally, the proposed speech processing pipeline using exclusively acoustic data performed competitively in comparison to the state-of-the-art acoustic as well as linguistic models to predict the disorder severity, offering better privacy and scope of generalized support for diversified users.

The newly curated dataset and the preliminary analysis on it have been published in the 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seoul, South Korea, April 14–20, 2024.

As the first author, I conducted 70% task in the experiments and documentation under the supervision of Prof. Eleni Stroulia.

A Machine-Learning Model for Detecting Depression, Anxiety, and Stress from Speech

Abstract: Predicting mental health conditions from speech has been widely explored in recent years. Most studies rely on a single sample from each subject to detect indicators of a particular disorder. These studies ignore two important facts: certain mental disorders tend to co-exist, and their severity tends to vary over time. This work introduces a longitudinal dataset labeled with depression, anxiety, and stress scores using the DASS-21 self-report questionnaire, and describes a machine-learning pipeline to determine the severity of the three mental disorders using acoustic features extracted from speech samples of this dataset. Our initial findings suggest that healthy participants adhere more to the study procedure than participants who exhibit indicators of depression, anxiety, and stress and demonstrate that a one-dimensional convolutional neural network, trained on VGG-19 features, predicts the severity of depression, anxiety, and stress with high accuracy.

5.1 Introduction

Mental disorders like depression, anxiety, and stress involve significant disturbances in thinking, emotional regulation, and behaviour, affecting the individual's day-to-day life and well-being. 12.5% of the world population lives with a mental disorder, among which depression and anxiety are most common[1]. In recent years, speech has been considered a reliable bio-signal for measuring the severity of high-priority mental disorders, including depression, anxiety, schizophrenia, stress, and Alzheimer's Dementia [102, 108, 109], because of its non-invasive nature and cost-effectiveness. Numerous machine learning models have been proposed by researchers to detect indicators of mental disorders [71, 110–112] etc. Although many of the mental disorders are interrelated and tend to co-exist, few studies propose prediction methodologies for comorbid conditions, such as depression, anxiety, and stress. The lack of datasets labelled with scores of multiple conditions is a major cause of the scarcity. The Distress Analysis Interview Corpus (DIAC-WoZ) is a well-known speech corpus labelled with depression and post-traumatic stress disorder (PTSD) [40]. Different

[1] https://www.who.int/news-room/fact-sheets/detail/mental-disorders

subsets of this dataset were introduced as the challenge corpus of the Audio-visual Emotion Recognition Challenge (AVEC) in 2016, 2017, and 2019 [16, 62, 94]. The DEPression and Anxiety Crowdsourced Corpus (DEPAC) dataset published in 2022 includes depression and anxiety labels on Patient Health Questionnaire–8 (PHQ-8) and General Anxiety Disorder-7 (GAD-7) scales, respectively [89]. Both of these datasets contain samples in the English language. To the best of our knowledge, there is no speech corpus labelled with depression, anxiety, and stress scores.

In this paper, we introduce a new longitudinal speech corpus containing over 1,000 speech samples in English and Spanish, collected from May 2022 to March 2023. The COVID-19 pandemic has increased the prevalence of mental health disorders, and studies show that during the pandemic, youth (15 to 39 years) were more vulnerable to depression and anxiety disorders [113]. Our dataset was collected during the post-pandemic period, recruiting participants between 19 and 29 years old, which will be a valuable resource for the researchers in this area. In our dataset, we have observed a significant positive correlation among the disorder scores, which supports the fact that these disorders tend to be comorbid. Our analysis also shows a positive correlation between disorder severity and participants' adherence to study protocol, indicating that healthy individuals demonstrate more conformity to routine activities than persons with mental disorders. Finally, we formulate a Convolutional Neural network (CNN) for predicting depression, anxiety, and stress scores on Depression Anxiety Stress Scales (DASS-21) with root-mean-square errors of 7.09, 7.69, and 8.40 out of 42. The model's performance is competitive with the state-of-the-art.

5.2 The Dataset

We collected speech samples from 40 participants between the ages of 19 to 29 years old from Mexico and Canada. 26 of the participants were native Spanish speakers (14 female, 9 male, and 3 identified as other gender), and 14 were English speakers (6 male, 8 female). Every three days, the participants provided two speech samples: guided reading and free-form speech. For guided reading, the participant read out the paragraph 'Please call Stella' [114] in English or Spanish. For the free-form speech task, the participants were asked two questions, randomly selected from a list of questions, including describing a memorable event, a hobby, or a favourite person. We developed an Android application for data collection. The application prompted the participants to record samples every three days. The data collection continued for two months, resulting in a corpus of 1,049 data points. 838 of the samples are in Spanish, and the rest are in English. Each audio sample ranges from 23 to 67 seconds in duration. We obtained 1 to 54 speech samples per participant, 26 on average.

After every recording session, participants are prompted to fill out the DASS-21 [115] questionnaire, consisting of 21 statements, with 7 questions associated with each of the scales of Depression, Anxiety, and Stress. The participants rated each statement from 0 to 3, indicating how much the statement applied to them. The three scales of the DASS-21 provide scores on depression, anxiety, and stress of the individual in the range of 0 to 21; these scores are then multiplied by 2 for consistency with the more detailed DASS-42 scale [115]. Individuals scoring lower than 9, 7, and 15 (out of 42) on the depression, anxiety, and stress scale respectively are considered healthy. 77%, 72%, and 88% of our samples belong to the normal range of depression, anxiety, and stress respectively. We summarized descriptive statistics of DASS-21 scores in our dataset in Table 5.1.

Table 5.1: Descriptive statistics the DASS-21 scores in each language.

		Depression	**Anxiety**	**Stress**
English	Mean	9.58	8.03	11.22
	Std.	8.29	8.00	8.71
	Min.	0.00	0.00	0.00
	Median	8.00	6.00	10.00
	Max.	36.00	32.00	34.00
Spanish	Mean	5.01	3.98	6.07
	Std.	7.07	6.36	7.74
	Min.	0.00	0.00	0.00
	Median	2.00	2.00	4.00
	Max.	42.00	42.00	42.00
Overall	Mean	5.94	4.79	7.12
	Std.	7.55	6.90	8.20
	Median	4.00	2.00	4.00

5.3 Predicting DASS-21 scores from Speech

We trained individual one-dimensional (1D) CNN models on VGG-19 features extracted from spectrograms of the audio samples to predict the severity of each disorder on DASS-21 scale.

5.3.1 Data Cleaning

We used the Noisereduce [2] algorithm to clean the speech samples. The algorithm computes the spectrogram of a speech signal and estimates a noise threshold for each frequency band of the signal. The threshold is used to compute a mask that filters the noise below the frequency-varying threshold [116].

[2] https://github.com/timsainb/noisereduce

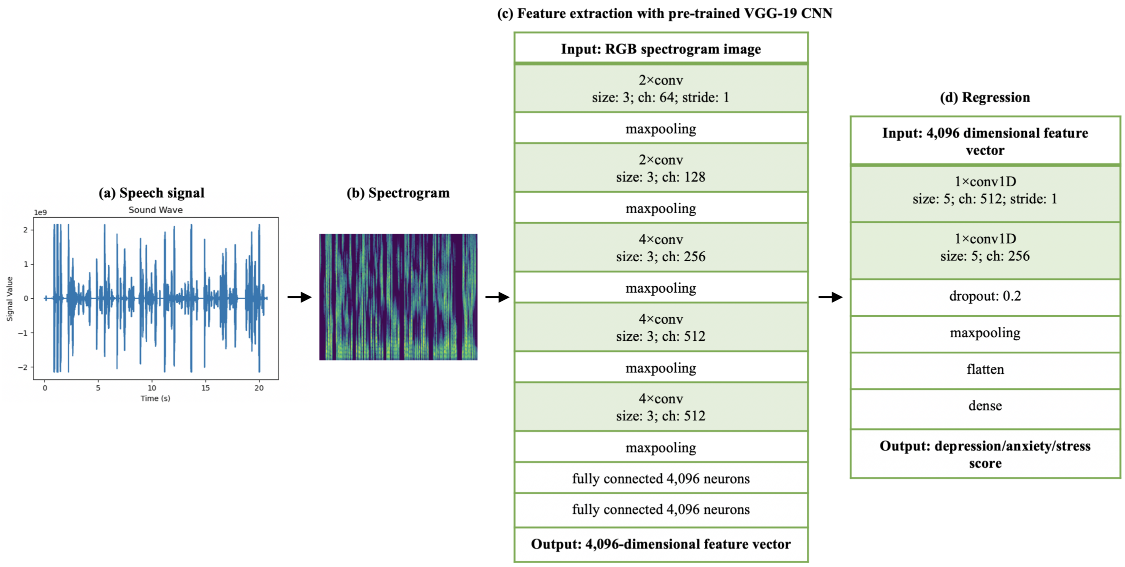

(c) Feature extraction with pre-trained VGG-19 CNN
Input: RGB spectrogram image
2×conv
size: 3; ch: 64; stride: 1
maxpooling
2×conv
size: 3; ch: 128
maxpooling
4×conv
size: 3; ch: 256
maxpooling
4×conv
size: 3; ch: 512
maxpooling
4×conv
size: 3; ch: 512
maxpooling
fully connected 4,096 neurons
fully connected 4,096 neurons
Output: 4,096-dimensional feature vector
(d) Regression
Input: 4,096 dimensional feature vector
1×conv1D
size: 5; ch: 512; stride: 1
1×conv1D
size: 5; ch: 256
dropout: 0.2
maxpooling
flatten
dense
Output: depression/anxiety/stress score
(a) Speech signal
Sound Wave
1e9
Signal Value
2
1
0
−1
−2
0
5
10
15
20
Time (s)
(b) Spectrogram

Figure 5.1: Our system pipeline. Spectrograms (b) are generated from whole audio files (a) and fed into pre-trained VGG-19 CNN. Activations of the second fully connected layer are extracted as 4,096-dimensional deep spectrum feature vectors (c) used to train a 1-dimensional CNN regressor (d).

5.3.2 Data Partitioning

We divided our data into five non-overlapping folds. In each fold, 20% of the samples were held out for testing. In each training and test partition, we ensured the same ratio of normal and high DASS-21 scores as the original dataset. Within each training fold, we used 20% samples for validation, maintaining the same proportion of normal and high scores as the original dataset. We also ensured that each speech sample appeared in at least one test set, and no speech sample appeared in multiple test sets.

5.3.3 Feature Extraction

We used the DeepSpectrum Python toolkit [3] for feature extraction from the audio samples with pre-trained CNNs. Hamming windows of width 16 ms shifting forward by 8 ms are used to compute the power spectral density on the dB power scale. Matplotlib[4] plots of 387×387 pixels in *viridis* colour map are generated, which are then resized to 224×224 pixels to fit the input size of CNN. *Viridis* is a sequential colour map varying from blue (lower range) to green to yellow (upper range) (Figure 5.1(b)).

The spectrograms are then fed into the pre-trained VGG-19 CNN [67]. VGG-19 CNN is a combination of 19 layers including convolutional, maxpooling, and fully connected layers, using rectified linear units (ReLU) as activation functions. To obtain the deep spectrum features, spectrogram plots are forwarded through the pre-trained

[3] https://github.com/DeepSpectrum/DeepSpectrum

[4] https://matplotlib.org/

networks, and the activations from the neurons on the second fully connected layers are extracted as feature vectors. The resulting feature set is a 4,096-dimensional vector, each representing one speech sample. Figure 5.1(c) illustrates the procedure of extracting VGG-19 features. Figure 5.1(a) to (d) depicts the complete pipeline of our system.

5.3.4 Experimental Setting

To predict depression, anxiety, and stress scores we formulated a CNN consisting of two 1D convolutional layers, followed by a dropout layer, a maxpooling layer, and a fully connected layer. ReLU is used as the activation function for the network layers. We used a filter size of 5 × 1 for the convolutional layers. The dropout rate was 0.2. The stride of the maxpooling was 8. ADAM optimizer was used, setting the learning rate to 10^{-5} with a decay of 10^{-7}. We trained the model for 300 epochs in batches of 32 samples. We applied early stopping when the root-mean-square-error (RMSE) loss did not decrease for 20 epochs. The prediction of depression, anxiety, or stress score was obtained from the final fully connected layer. After training on each fold, we obtained the prediction on the test fold. As the test folds collectively contain all the speech samples in our dataset, we report the model performances on the concatenated predictions on the five test folds.

5.4 Results and discussion

We report the performance of our model using the following two metrics:

$$\mathrm{RMSE}(y,\hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}} \tag{5.1}$$

Here y_i and $\hat{y}_i$ represent the ground truth and predicted scores on *ith* sample, and N indicates the total number of samples.

$$R^2(y,\hat{y}) = 1 - \frac{(y_i - \hat{y}_i)^2}{(y_i - \overline{y})^2} \tag{5.2}$$

where $\overline{y}$ is the mean of the ground truth scores. Table 5.2 summarized the performance of our regression model in predicting the score of the three disorders.

In our work, we predicted DASS-21 scores for each disorder, ranging from 0 to 42. As the scores in existing literature report their predictions in different scales, we compare our performance with the state-of-the-art using the normalized root-mean-square-error (NRMSE) metric calculated as follows:

$$\mathrm{NRMSE} = \frac{RMSE}{y_{max} - y_{min}}\% \tag{5.3}$$

Here y_{max} and y_{min} are the highest and lowest values on the measurement scale respectively. Table 5.3 shows the comparison of regression models in existing literature with our proposed model.

Table 5.2: RMSE and R^2 metrics of the DASS-21 predictions of our proposed CNN model

	RMSE	R^2
Depression	7.09	0.37
Anxiety	7.69	0.47
Stress	8.40	0.36

Table 5.3: Comparison of regression models for predicting depression, anxiety, and stress. (*) indicates multimodal model.

	Citation	**Scale**	**Range**	**NRMSE**
Depression	Kim *et al.*[117]	PHQ-9	27	0.205
	Rodrigues *et al.* [59]	PHQ-9	27	0.211
	He & Cao [23]	BDI-II	63	0.159
	Tasnim & Stroulia [71]	BDI-II	63	0.155
	Tasnim *et al.* [118]	PHQ-8	24	0.221
	Ray *et al.* [58]	PHQ-8	24	0.213
	Our Study	DASS-21	42	0.169
Anxiety	Fatima *et al.** [119]	DASS-21	42	0.089
	Our Study	DASS-21	42	0.183
Stress	Fatima *et al.** [119]	DASS-21	42	0.103
	Our Study	DASS-21	42	0.200

Table 5.3 shows that our CNN model outperforms most other speech-based models for predicting depression severity and is competitive on the other two disorders, in two languages. There are no other studies predicting anxiety and stress based on acoustic features only. Fatima *et al.* [119] used linguistic features extracted from text data. Our CNN models using acoustic features exclusively predict anxiety and stress with a competitive error ratio. One limitation of our dataset is that most of our samples fall in the normal range of depression, anxiety, and stress scores respectively, which biases the model's prediction towards subthreshold scores. In our future work, we will consider balancing the dataset by augmenting samples within the higher range of scores. In this work, we considered each sample as an independent instance, as each sample is associated with a DASS-21 score. We plan to explore the possibilities of formulating personalized models by exploiting the longitudinal dataset in our future endeavours.

5.5 Conclusion

In this work, we introduce a new longitudinal and multilingual (English and Spanish) speech corpus for depression, anxiety, and stress. Our dataset captures valuable information on the post-pandemic effect on the mental health of youths. The dataset supports the fact that individuals with lower levels of depression, anxiety, and stress exhibit more conformity with routine activities, demonstrated by a positive correlation between DASS-21 scores and adherence to our data collection protocol. Finally, we propose a CNN model trained on VGG-19 features extracted exclusively from acoustic data recorded in English and Spanish language. In comparison to the state-of-the-art acoustic as well as linguistic models, our proposed model demonstrates competitive performance. The usage of acoustic features exclusively offers two benefits. Firstly, the data does not require going through any transcription, therefore it ensures better privacy of the content of the speech. Secondly, being language-independent in nature, this kind of model extends support to diversified users.

Chapter 6

Learning Language and Acoustic Models for Identifying Alzheimer's Dementia from Speech

The speech-processing pipeline described in this chapter was developed to combine acoustic and linguistic aspects of speech to determine the prevalence and severity of AD. As part of a multidisciplinary research group, I implemented a version of the depression prediction pipeline, adapted to the task of predicting AD vs healthy individuals and measuring the severity on the MMSE scale. The dataset was published as the ADReSS Challenge 2020 corpus.

- **Data Preprocessing:** The speech samples were segmented using timestamps provided in the transcripts to remove the interviewers' utterances. Three acoustic feature sets, i.e., AVEC 2013 baseline, ComParE, and emo_large, were extracted from the segments to train several classifiers and regressors. These feature sets were composed of low-level descriptors (LLDs) consisting of energy, spectral, MFCC, and voicing-related features and statistical and regressional functionals computed on the LLDs. The paper also described models on linguistic features explored by other members of the group.

- **Model Training:** The learning algorithms explored in this experiment are logistic regression, random forest, SVM, and XGBoost. The individual classifiers and regressors were trained and optimized using a 5-fold internal cross-validation. Accuracy and regression error were computed by averaging across (external) 5 folds. The models with the highest accuracy and lowest regression error were combined to formulate two ensemble models: one based on acoustic features only and another based on acoustic and linguistic features.

- **Prediction:** Weights were assigned to each classification model proportionate to that model's mean cross-validation accuracy. Then a linear weighted combination of the classifiers' predictions was computed to obtain the ensemble prediction. An unweighted average of the predictions made by the best models was considered as predicted MMSE scores.

Key findings: In this work, multiple combinations of features and ML algorithms were explored. The key findings can be summarized as follows:

1. The weighted-majority-vote ensemble of the best model on linguistic features and three of the best models on acoustic features obtained the highest average cross-validation accuracy (81% ± 1.17%).

2. The random forest model obtained the best RMSE of 5.62 on the held-out set. The ensemble of random forest on linguistic features and gradient boosting tree using acoustic features did not perform as well as the standalone random forest model on linguistic features, with an RMSE of 6.12 on the test set. The linguistic features, such as semantics, fluency, and n-grams were found to be the most useful.

3. The results, of over 80% accuracy for classification and approximately 6.00 RMSE for regression, demonstrate the promise of using speech-based ML models to detect cognitive decline from speech.

The analysis done in this chapter was published in the Frontiers in Computer Science 3 (2021): 624659, describing our methodology for competing in the ADReSS 2020 challenge.

As a participating team member, I worked on the acoustic pipeline, which is approximately 30% of the experimentation. I also compiled about 30% of the documentation.

Learning Language and Acoustic Models for Identifying Alzheimer's Dementia from Speech

Alzheimer's disease (AD) is a chronic neurodegenerative illness that manifests in a gradual decline of cognitive function. Early identification of AD is essential for managing the ensuing cognitive deficits, which may lead to a better prognostic outcome. Speech data can serve as a window into cognitive functioning and can be used to screen for early signs of AD. This paper describes methods for learning models using speech samples from the DementiaBank database, for identifying which subjects have Alzheimer's dementia. We consider two machine learning tasks: (a) binary classification to distinguish patients from healthy controls, and (b) regression to estimate each subject's Mini-Mental State Examination (MMSE) score. We develop models based on acoustic and language features and explore a variety of dimension-reduction techniques, training algorithms, and fusion strategies. Our best-performing classification model, using language features with dimension reduction and regularized logistic regression, achieves an accuracy of 85.4% on a held-out test set. On the regression task, a linear regression model trained on a reduced set of language features achieves an RMSE of 5.62 on the test set. Our initial results demonstrate the promise of using machine learning for detecting cognitive decline from speech in AD patients.

6.1 Introduction

Alzheimer's Dementia (AD) has recently become one of the leading causes of death in people over 70 years [120]. With life expectancy increasing, the prevalence of AD among older adults is also rising. Currently, the number of cases among people over the age of 60 is doubling every 4-5 years, and one in every three individuals over the age of 80 is likely to develop AD [121]. AD is a progressive neurodegenerative disorder that is characterized by the loss of subcortical neurons and synapses that begins in areas such as the hippocampus and the entorhinal cortex [122, 123] Over time, more associative areas begin to show amyloid deposition and neurofibrillary tangles in addition to neuronal and synaptic loss. As it spreads, patients develop additional cognitive and functional deficits in domains such as attention, executive function, memory and language [124]. Current theories maintain that clinical symptoms are preceded by subtle cognitive deficits that worsen over time. Early recognition of these deficits could prove valuable for treating pre-stage AD, allowing for a better quality of life for the patient and their caregivers.

Currently, clinical diagnostic methods for determining who has AD include cognitive assessments (e.g., Mini-Mental State Examination [MMSE]), self-report questionnaires and neuroimaging (e.g., Positron Emission Tomography [PET]) [125]. While these methods have proven useful, they suffer from several shortcomings. Cognitive assessments can be tedious and suffer from test-retest reliability or practice effects; self-report questionnaires also lack reliability and validity; and neuroimaging is an expensive, invasive, and time-consuming procedure. A simple, non-invasive and inexpensive approach such as speech data could be useful for detecting AD. Episodic memory, visuospatial ability, and confusion are some of the first signs of cognitive decline in AD patients [126, 127]. These deficits can be observed through verbal communication in a structured task, motivating the recent use of speech data for diagnostic screening of AD in elder patients [128]. In our study, we used machine learning (ML) approaches to distinguish between AD and control patients using acoustic and linguistic features from spontaneous speech during a picture description task.

The current literature on detecting AD from spontaneous speech samples can be divided into two main categories. One class of systems analyzes linguistic features (lexicon, syntactic and semantic information) while the other deals with acoustic-dependent features. In the acoustic domain, AD patients exhibit longer and more frequent hesitations, lower speech and articulation rates, and longer pauses compared to control participants in spontaneous speech tasks [129, 130]. Some have attempted to apply ML approaches to learn models that use acoustic features to distinguish AD from control participants. Toth *et al.* learned a model for classifying early-stage AD patients from control patients using spontaneous speech from a recall task [131]. Their classification model (Random Forest) achieved an F1-score of 78.8%, and they found significant differences in speech tempo, articulation rate, silent pause, and length of utterance. Mirzaei *et al.* tried to improve on previous models by examining temporal features (jitter, shimmer, harmonics-to-noise ratio, Mel frequency cepstral coefficients [MFCCs]) [132]. Using a two-stage feature selection process, they were able to improve on previously reported model accuracies by 30% in distinguishing early-stage AD from control patients.

Conversational transcripts contain rich information about the speaker, such as the health of their vocabulary, the complexity of their syntactic structures, and the information and meanings they communicate. Previous research has shown that language changes in patients who suffer from AD [133, 134] – e.g., these patients often have difficulty naming objects within specific categories, replacing forgotten words with pronouns and repeating certain words or phrases [135–137]. This has motivated numerous research projects on conversation samples in AD and control patients. Fraser *et al.* examined picture description transcripts from demented and control individuals [138].

Subsequently, they also analyzed acoustic features in addition to natural language and achieved an accuracy of 81%. They found that semantic information was one of the best features (syntactic fluency, MFCCs and phonation rate were also among the top features) for separating AD from control patients. Our study hopes to improve further by utilizing different natural language and acoustic processing approaches.

In particular, we will train models that use both acoustic and context-dependent features to distinguish AD from healthy elders. A secondary goal is to use these same features to assess symptom severity from the MMSE test scores of these patients. We utilize both acoustic and linguistic features from short speech samples to determine whether someone may be suffering from AD or not. These separate features are fed into different pipelines (acoustic and linguistic) where they will be pre-processed, and then cross-validation will be used to tune hyper-parameters and select features. Afterward, we explore ways to combine the various models using ensemble methods to produce models that can label a speech sample as either AD or non-AD, and also to predict their MMSE scores.

The rest of the paper is organized as follows. Section 2 provides an overview of our materials and methods; Section 3 outlines the feature sets we extract from the data and the algorithms we use to construct our classification and regression models; Section 4 reports on our results; and Section 5 concludes with a statement of our contributions and plans for future work.

6.2 Method

For this study, we were given a training set of 54 AD subjects and an age- and gender-matched set of 54 healthy controls. For each subject, we obtained the original recorded speech sample, normalized speech segments extracted from the full sample using voice activity detection and noise removal, as well as speech transcript files annotated using the CHAT (Codes for Human Analysis of Transcripts) transcription format [139]. Additionally, some descriptive features were given about these individuals, including age, gender, binary class label (AD/non-AD), and their MMSE score. The challenge organizers withheld a test set containing 24 AD and 24 healthy subjects for final evaluation.

We used 5-fold cross-validation to evaluate our trained models. To ensure consistent and reliable comparison between our models, we defined and used a common set of folds that were balanced in terms of class labels (or MMSE scores) as well as gender. For each model, performance metrics are reported (average accuracy, average RMSE, etc.) based on these held-out folds, as well as on the final hold-out test set (where available).

6.2.1 Language and fluency features

The organizers provided transcripts that were annotated using the CHAT coding system [139]. Using the CLAN (Computerized Language ANalysis) program for processing transcripts in the CHAT format, we computed the following set of global syntactic and semantic features for each transcript: type-token ratio (TTR) – the number of unique words divided by total number of words; mean length of utterance (MLU), where an utterance is a speech fragment beginning and ending with a clear pause; number of verbs per utterance; percentage of occurrence of various parts of speech (nouns, verbs, conjunctions, etc.); number of retracing (self-corrections or changes); and number of repetitions. We also computed a number of fluency features, including percent of broken words, part-word and whole-word repetitions, sound prolongations, abandoned word choices, word and phrase repetitions, filled pauses, and non-filled pauses. In total, we computed 62 such informative summary features from the transcripts.

6.2.2 N-gram features

We processed the raw (unannotated) transcripts to compute bag-of-words and bigram features. First, we standardized the transcripts by converting them into lists of word tokens. Next, we used the WordNet lemmatizer to find and replace each word with the corresponding lemma; for example, words like "stands", "standing" and "stood" were all replaced by the common root word "stand". Finally, we removed stopwords from each transcript, where stopwords are highly common (and, we assume, uninformative) words that may add noise to the data (such as "I", "am", "was", etc.), using a predefined stopwords list from the Python natural language toolkit (NLTK) package. After these preprocessing steps, the transcripts were ready for computing n-gram features.

Next, we used the standardized transcripts to compute bag-of-words vectors (using words seen in the training set only) and normalized these vectors with the Term Frequency-Inverse Document Frequency (TF-IDF) function. Bag-of-words is a common language representation, where the frequency of word occurrences is used to build a feature vector (no context information). TF-IDF can be viewed as a normalization procedure as it reflects how important a word is to a document in a corpus – effectively penalizing words that occur frequently in most of the documents in the corpus. For example, in our case the word "boy" might occur frequently in all transcripts, so it may not be very informative. Finally, we also computed bigram vectors in a manner similar to bag-of-words.

6.2.3 Acoustic features

Using the speaker timing information provided in the transcripts, we extracted the participants' utterances (removing the clinician's voice) from the audio recordings, for a total of 1,501 participant utterances from the training set, and 592 from the test set. We then normalized the audio volume across all speech segments. Four different sets of features were manufactured from each audio segment using OpenSMILE v2.1 [85]. Note that our overall learner will consider various base-learners, each running on one of these feature sets.

(FeatureSet#1) The **AVEC 2013** [14] feature set includes 2,268 acoustic features including 76 low level descriptor (LLD) features and their statistical, regression and local minima/maxima related functionals. The LLD features include energy, spectral and voicing related features; delta coefficients of the energy/spectral features, delta coefficients of the voicing related LLDs and voiced/unvoiced duration based features.

(FeatureSet#2) The **ComParE 2013** [72] feature set includes energy, spectral, MFCC, and voicing related features, logarithmic harmonic-to-noise ratio (HNR), voice quality features, Viterbi smoothing for F0, spectral harmonicity and psychoacoustic spectral sharpness. Statistical functionals are also computed, leading to a total of 6,373 features.

(FeatureSet#3) Our third feature set consists of the following three feature sets. The **emo_large** [85] feature set consists of cepstral, spectral, energy and voicing related features, their first and second order delta coefficients as LLDs; and their 39 statistical functionals. The functionals are computed over 20 ms frames in spoken utterances. This produced 6552 acoustic features across the utterances. The **Jitter-shimmer** feature set is a subset of INTERSPEECH 2010 Paralinguistic Challenge [69] feature set, consisting of 3 pitch related LLDs and their delta coefficients. We also computed 19 statistical functionals of the LLDs on the voiced sections of the utterances, resulting in 114 features. Finally, we extracted 7 speech and articulation rate features by automatically detecting syllable nuclei [140], and use a script from the software program Praat to detect peaks in intensities (dB) followed by sharp dips. We also calculated other features, such as words per minute, number of syllables, phonation time, articulation rate, speech duration and number of pauses for each speech sample [141].

(FeatureSet#4) We computed the **MFCC 1-16** features and their delta coefficients from 26 Mel-bands,which uses the fast Fourier transform (FFT) power spectrum. The frequency range of the Mel-spectrum is set from 0 to 8 kHz. Inclusion of statistical functionals resulted in 592 features. This feature set is a subset of AVEC 2013 feature set [14].

We also added age and gender of the participants to each set of features. In our audio based model, FeatureSet#1 was used only for the classification task while FeatureSet#2 was used only for the regression task. FeatureSets#3 and #4 were used for both classification and regression tasks.

6.2.4 Language-based models

Given our two sets of linguistic features above (Sections 6.2.1 and 6.2.2), we explored various dimension reduction techniques and learning algorithms to find the best performing pipeline. The dimension reduction techniques include Principal Component Analysis (PCA), Latent Semantic Analysis (LSA), and univariate feature selection using ANOVA F-values. The learning algorithms explored for the classification task are logistic regression (LR), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGB). For the regression task, the regression versions of the same algorithms are trained (except logistic regression is replaced by linear regression). Internal 5-fold cross-validation was used to tune the hyperparameters for these models.

Our internal cross-validation found the best-performing language-based classification model, which consisted of the following steps:
Step1: 5-component PCA transformation of the language and fluency features (after standardizing using z-scores);

Step2: 50-component LSA transformation of the N-gram features (after standardizing using TF-IDF transform); and
Step3: L1-regularized logistic regression

The best language-based regression model involved the following:
Step1: 30-component PCA transformation of the language and fluency features (after standardizing using z-scores);
Step2: 100-component LSA transformation of the N-gram features (after standardizing using TF-IDF transform); and
Step3: Random Forest Regressor, using 100 trees, minimum of 4 instances at each leaf node, and 25 features considered for each split.

6.2.5 Acoustic models

All acoustic features were real values and were therefore standardized using z-scores – i.e., subtracting the mean and dividing by standard deviation. We used PCA to reduce the dimensionality of the features sets. For FeatureSet#1 and FeatureSet#2, PCA was performed to identify the minimum number of features capable of retaining 95% of the variance. In case of FeatureSet#3 and FeatureSet#4 the number of principals were determined through internal 5 fold cross validation. Therefore, the dimension of FeatureSet#1 is reduced to 700, FeatureSet#2 to 1100, FeatureSet#3 to 1000 and FeatureSet#4 to 50. Next, we selected the best 50 principal components from FeatureSet#1, and the best 70 from FeatureSet#3 applying univariate feature selection method based on ANOVA F-value between label and feature. For FeatureSet#2, we calculated feature importance weights using a decision-tree regression model, and selected only the features with importance weight higher than the mean.

After this pre-processing stage, our system fed these audio features to various machine-learning algorithms, that each identify patterns of features that can distinguish dementia patients from healthy controls (the classification task), and can compute a subject's MMSE score (the regression task). We explored several learning algorithms, including Adaboost, XGB, RF, gradient boosting (GBT), decision tree (DT), hidden Markov model (HMM) and neural network (NN). Internal 5-fold cross-validation was performed to tune the hyperparameters of the classifiers and regressors. The predictions were made in two steps following the challenge baseline. In the first step, the classifiers and regressors were trained and tested with acoustic features, age and gender to predict whether the speech segment was uttered by a health control or an AD patient and to predict that subject's MMSE score. Next, weighted majority vote classification was performed to assign each subject a label of health control or AD, based on the majority labels of the segment-level classification. The predicted MMSE scores on all the segments of one subject were averaged to calculate the final MMSE score of that subject. The best-performing classifiers on acoustic data are the following:
(1) Neural network with 1 hidden layer, trained on FeatureSet#1
(2) AdaBoost Classifier with 50 estimator and logistic regression as base estimator, trained on FeatureSet#4
(3) Adaboost with 100 estimators and DT as the base estimator trained on Feature-

Table 6.1: Classification model results

Classifiers	**Class**	**Precision**	**Recall**	**F1 Score**	**Accuracy**	**Accuracy (Hold-out set)**
Logistic Regression (NLP)	AD	0.71	0.60	0.75		
	HC	1.00	1.00	0.83		
	OVR	0.80	0.80	0.79	**80% ± 0.00%**	**85%**
SVM (NLP)	AD	0.68	0.84	0.75		
	HC	0.79	0.60	0.68		
	OVR	0.73	0.72	0.72	72% ± 1.85%	73%
Majority vote (NLP + Acoustic)	AD	0.74	0.96	0.83		
	HC	0.94	0.66	0.78		
	OVR	0.84	0.81	0.81	**81% ± 1.17%**	**83%**
Majority vote (Acoustic)	AD	0.71	0.78	0.74		
	HC	0.76	0.68	0.72		
	OVR	0.73	0.73	0.73	73% ± 1.36%	65%
Baseline (Acoustic)	AD	0.57	0.52	0.54		
	HC	0.56	0.61	0.58		
	OVR	0.57	0.57	0.56	57%	63%

AD Alzheimer's dementia   HC Healthy control   OVR Overall rating

Set#3.

The three regressors with the lowest root mean square error (RMSE) were
(1) Gradient boosting regressor, trained on FeatureSet#4
(2) Decision tree with number of leaves 20, trained on FeatureSet#2
(3) Adaboost regressor trained on FeatureSet#3 with 100 estimators.

6.2.6 Ensemble methods

After obtaining our best-performing acoustic and language-based models, we computed a weighted majority-vote ensemble meta-algorithm for classification. We chose the three best-performing acoustic models along with the best-performing language model and computed a final prediction by taking a linear weighted combination of the individual model predictions. The weights assigned to each model were proportional to that model's mean cross-validation accuracy. For regression, we also computed an unweighted averaging of our best language and acoustic model predictions for MMSE scores.

6.2.7 Results

6.2.8 Classification

Table 6.1 presents the results for the classification task. The model that obtained the highest average cross-validation accuracy (81% ± 1.17%) is a weighted-majority-vote ensemble of the best language-based model and three of the best acoustic-based models. The second highest accuracy (80% ± 0.00%) was obtained by the language-based logistic regression. However, a t-test reveals that these two models do not exhibit a statistically significant difference in performance ($t_{(4)} = 0.34$, $P > 0.05$). This is also evident by the performance of these two models on the final held-out set, where the language-based logistic regression gives the highest accuracy (85%) and the weighted-majority-vote ensemble gives a slightly lower accuracy (83%).

Note that our ensemble model, which uses only acoustic features, performs significantly better than the "baseline model" (provided by the organizers), which also uses acoustic features only.

6.2.9 MMSE prediction

Table 6.2 shows the root-mean-square error (RMSE) of various regression models; column 2 shows the average RMSE score over the 5 cross-validation folds, and column 3, on the hold-out test set (provided by the organizers of the challenge). These results show that the language-based model obtains the best RMSE of 6.43 on the cross-validation set and 5.62 on the hold-out set. The combined language-acoustic model did not perform as well as the standalone language-based model, with an average RMSE of 6.83 on the cross-validation set and 6.12 on the hold-out set. However, a t-test between these RMSEs of the two best models (best acoustic + best language-based combination, vs. best standalone language-based), shows they have similar performance ($t_{(4)} = 0.25$, $P > 0.05$).

6.3 Conclusion

We investigated a variety of ML models, using language and acoustic features, to identify models that performed well at using speech information to distinguish AD from healthy subjects, and to estimate the severity of AD. Our results, of over 80% accuracy for classification and approximately 6.00 RMSE for regression, demonstrate the promise of using ML for detecting cognitive decline from speech. In our investigation, we explored multiple different combinations of features and ML algorithms; in the future, it would be interesting to delve deeper into the behaviour of our best models, to determine the contribution of individual (or groups of) features to the model's ability to distinguish AD patients from healthy controls. Further, although we have currently used the full set of standard stopwords for removing noise in our language models, it may be worthwhile to see whether using a reduced set of stopwords (for example, not removing pronouns) might be more advantageous.

Table 6.2: Regression model results.

Regressors	**RMSE**	**RMSE (Hold-out Set)**
Random Forest (NLP)	6.43 ± 0.18	**5.62**
Gradient Boosting (Acoustic)	6.89 ± 0.17	6.67
Random Forest (NLP) + Gradient Boosting (Acoustic)	6.66 ± 0.18	**6.01**
Majority vote (All models)	6.85 ± 0.16	6.12
Baseline (Acoustic)	7.30	6.14

Our current best-performing models outperform recent results reported in the literature and provide evidence that for discriminating between subjects with AD versus healthy controls, features based on language (semantics, fluency, and n-grams) are very useful. Furthermore, a weighted majority vote of acoustic and language-based models demonstrates competitive performance, implying that a combination of acoustic and language features also holds potential. Finally, comparing only acoustic models, we find that accuracy improves significantly compared to the baseline model for both the classification and regression tasks.

Chapter 7

Machine-Learning Models for Detecting Mild Cognitive Impairment in Multilingual Speech

In the most recent experiment described in this chapter, I applied the speech-processing pipeline to TAUKADIAL 2024 MCI detection challenge tasks, as part of a collaborative research group. The challenge task was to predict the binary class (MCI vs. healthy) and MCI severity of subjects on the MMSE scale.

- **Data Preprocessing:** At the beginning of the experiment, the language of the samples (Chinese or English) was identified, which was used as a feature while training. Disfluency features including total length of each sequence, average duration of sequences, duration of the longest sequence, and count of sequences were calculated for the participants' segments, the interviewers' segments, and the periods of silence. Then, for the subjects' segments only, 500-dimensional Bag-of-Audio-Words features were extracted using ComParE low-level acoustic descriptors. The disfluency features, BoAW features, and demographic attributes of the participant, i.e., age and sex, and language identification (English or Chinese) were horizontally concatenated for each speech sample. Collectively, this feature set contained information on the participants' acoustic fluency, disfluency, and demographic information relevant to their cognitive health. The other members of the group explored pre-trained acoustic embeddings (VGGish and Wav2Vec2), pre-trained semantic embeddings (OpenAI's `text-embedding-3-large`), and a zero-shot learning framework using GPT-4 API to analyze the transcripts from audio recording.

- **Model Training and Prediction:** Several machine learning models were trained in this experiment including logistic regression, SVM, random forest, XGBoost, and neural network. The hyperparameters of these models were tuned using 5 subject independent internal cross-validation folds on the training data. The following base models were formulated to perform the two tasks specified in the challenge:

Base Classifiers:

Semantic classifier: Logistic regression (LR) model trained on GPT4 features

Fluency classifier: Neural network (NN) model trained on a combination of disfluency features and GPT4 features

Acoustic classifier #1: Random forest (RF) model trained on wav2vec features.

Acoustic classifier #2: RF model trained on VGGish features

Base Regressors:

Semantic regressor: Linear regressor trained on GPT4 features following PCA dimension reduction

Fluency regressor: XGBoost model trained on a combination of acoustic disfluency features, word disfluency features, and GPT4 features

Acoustic regressor #1: SVM model trained on VGGish features

Acoustic regressor #2: RF model trained on VGGish features

To leverage the strengths of the individual base models, the semantic, acoustic, and fluency models' predictions were ensembled by averaging the classification prediction probabilities and regression values to formulate the following ensemble models:

Ensemble classifiers:

Ensemble classifier #1: SVM classifier trained on concatenated disfluency, BoAW and demographic features

Ensemble classifier #2: Combination of *semantic classifier*, *fluency classifier* and *acoustic classifier #1*

Ensemble classifier #3: Combination of *semantic classifier*, *fluency classifier* and *acoustic classifier #2*

Ensemble regressors:

Ensemble regressor #1: Random forest regressor trained on concatenated disfluency, BoAW and demographic features

Ensemble regressor #2: Combination of *semantic regressor*, *fluency classifier* and *acoustic regressor #1*

Ensemble regressor #3: Combination of *semantic regressor*, *fluency classifier* and *acoustic regressor #2*

- **Prediction:** The *fluency classifier* and the *Ensemble regressor #1* provided the highest classification accuracy and lowest regression error, respectively, on the test partition of the dataset.

Table 7.1: Classification performance of the models. The scores are averaged across 5-folds (20-folds for the baseline). The Rightmost column represents UAR score on the test set. The best UAR scores are marked in bold

Model	**Sensitivity**	**Specificity**	F_1	**UAR (5-fold)**	**UAR (test)**
Semantic classifier	0.784	0.618	0.758	0.702	0.568
Fluency classifier	0.821	0.546	0.759	0.683	**0.570**
Ensemble Classifier #1	0.685	0.558	0.621	0.622	0.515
Ensemble Classifier #2	0.891	0.618	0.820	**0.754**	0.563
Ensemble Classifier #3	0.891	0.600	0.815	0.745	0.515
Baseline [142]	-	-	-	0.509	0.592

Table 7.2: Regression RMSE of the models. The lowest RMSE scores are marked in bold

Model	RMSE (5-folds)	RMSE (test)
Semantic regressor	2.72	3.21
Fluency regressor	2.74	3.69
Ensemble regressor#1	2.66	**2.54**
Ensemble regressor#2	**2.56**	3.23
Ensemble regressor#3	2.62	3.26
Baseline [142]	2.86	2.89

Key findings:

- The ensemble classifier exploiting the late-fusion approach yielded achieving 75.4% UAR in 5-fold CV which was significantly higher than the baseline UAR and performed competitively on the test set. The fluency model performed marginally better than the ensemble classifier on the test partition Table 7.1.
- The adaptation of the speech processing pipeline for MCI detection, exploiting acoustic fluency and disfluency features, coupled with relevant demographic and language identification, performed significantly better on the regression task (Table 7.2) than the challenge baseline.

The experimental methodology to address the TAUKADIAL 2024 challenge has been submitted to INTERSPEECH 2024, Kos Island, Greece.

As a member of the participating team and the first author of the paper, I conducted approximately 40% of the experimentation and 70% of the documentation.

Machine-Learning Models for Detecting Mild Cognitive Impairment in Multilingual Speech

In this paper, we describe our methods for learning models to perform two tasks on the TAUKADIAL 2024 English and Chinese speech samples: binary classification, to distinguish MCI patients from healthy controls, and regression to estimate each subject's MMSE score. We pursued two different methodologies. As an early-fusion approach, we concatenated acoustic, dysfluency, demographic, and language identification features, and we trained an SVM and a random forest model on this feature set. In a late-fusion approach, we trained three base models on text embeddings, disfluency features, and wav2vec acoustic features, and averaged the predicted probabilities and regression scores of these three models. The early-fusion model obtained an average classification accuracy of 62.60% and a regression RMSE of 2.66 on a 5-fold CV. In contrast, the late-fusion model yielded a classification accuracy of 77.54% and an average regression RMSE of 2.56.

7.1 Introduction

Mild Cognitive Impairment (MCI) is a neurological condition characterized by cognitive decline that is beyond typical age-related changes but does not significantly hinder daily functioning. Approximately 10-20% of individuals with MCI, aged 65 and older, are likely to progress to dementia.

In this paper, we describe our methodologies for learning models to identify which subjects have MCI, from the TAUKADIAL 2024 English and Chinese speech – training on both languages, to produce a model that tested on both. We consider two tasks: binary classification, to distinguish MCI patients from healthy controls, and regression, to estimate each subject's Mini-Mental State Examination (MMSE) score.

We experiment with a variety of acoustic and linguistic features, employing several training algorithms and fusion strategies. As an early-fusion approach, we concatenate Bag-of-Audio-words (BoAW) features derived from low-level speech descriptors, disfluency features, demographic attributes (i.e., age and sex), and language identification. We train a support vector machine (SVM) and a random forest model on the concatenated feature set for classification and regression tasks, respectively. For the late-fusion approach, we first train three base models on distinct feature sets, i.e., text embeddings from transcripts, disfluency features, and wav2vec acoustic features. The late-fusion classification and regression predictions were derived by averaging the predicted probabilities and regression scores of these three single-modality models.

We evaluated the models' performance using 5-fold cross-validation (5-fold CV). The early-fusion model obtained an average classification accuracy of 62.60% (SD=4.34%) and regression root-mean-square-error (RMSE) of 2.66 (SD=0.37) on the five folds. In contrast, the late-fusion model yielded a classification accuracy of 77.54% (SD=2.73%) and an average regression RMSE of 2.56 (SD=0.58).

On the held-out test set, the early-fusion model attained a classification accuracy of 51.66%, while the late-fusion model outperformed it with a 57.5% accuracy. In terms of regression, the early-fusion model achieved an RMSE of 2.54, whereas the late-fusion model scored 3.23 on the test set.

These initial results demonstrate the promise of using language-agnostic machine-learning models for detecting cognitive decline from speech for early intervention and management of MCI.

7.2 Background

In recent years, there has been significant research activity around using speech and language to infer the severity and progression of many mental illnesses and neurodegenerative disorders including mild cognitive impairment and Alzheimer's dementia (AD). These disease states produce measurable changes in a variety of speech parameters. For example, Alzheimer's and dementia patients are prone to use wrong and meaningless words, and their pause duration is increased [143]. A variety of speech processing techniques, acoustic and linguistic features, feature selection algorithms, and machine learning models have been explored by researchers to identify the most effective combination for the task.

Fraser *et al.* [144] presented an analysis of Cookie Theft narratives in English and Swedish to detect mild cognitive impairment. They generated multilingual topics by clustering multilingual word embedding. Their proposed multilingual topic models outperformed monolingual models in both languages, achieving classification accuracy of 63% (English) and 72% (Swedish). Themistocleous *et al.* proposed a Deep Neural Network architecture, that aims to identify MCI from Swedish speech [145]. The classifier trained on acoustic features including vowel duration, vowel formants (F1 to F5), and fundamental frequency demonstrated 83% accuracy (SD=15%) in 5-fold cross-validation.

Wang *et al.* [146] aimed at detecting individuals with MCI among Chinese-speaking elderly individuals using multiple spoken tasks and uncovered task-specific contributions with a tentative interpretation of features. Their late-fusion configuration of task-specific models demonstrated an F_1 score of 0.96 in classifying MCI patients and outperformed each task-specific model.

Calzà *et al.* [147] presented spoken language analysis applying Natural Language Processing (NLP) techniques to identify minor language alterations in potential MCI patients. Their proposed SVM classifier trained on a combination of acoustic, rhythmic, lexical, syntactic, and readability features extracted from an Italian language corpus could achieve F_1 score of 74.5%.

Vincze *et al.* [148] exploited linguistic features like characteristics of spontaneous speech, morphological and syntactic parsing for training an SVM classifier to identify Hungarian patients suffering from MCI achieving F_1 score of 69.1% which improved to 75% by applying feature selection based on statistical significance.

Asgaria *et al.* [149] grouped spoken words using Linguistic Inquiry and Word Count to categorize 2500 English words into 68 different word subcategories such as positive and negative words, fillers, and physical states. Their support vector classifiers distinguished MCI from cognitively intact participants with 85% accuracy.

Toth *et al.* developed a model for distinguishing early-stage AD patients from control patients using spontaneous speech from a recall task [131]. Their Random Forest classifier achieved an F_1 score of 78.8%, and they found significant differences in speech tempo, articulation rate, silent pause, and length of utterance between speech samples of AD patients and controls.

The 2023 ADReSS-M Signal Processing Grand Challenge aimed to explore transferable and generalizable speech features across languages for AD prediction by defining a prediction task where participants trained their models on English speech data and assessed their models' performance on spoken Greek data. A variety of acoustic features, as well as dysfluency and pause features derived from automatic speech recognition (ASR) were found effective by the top-ranking challenge participants [150–153]

7.3 Materials and Method

7.3.1 Dataset

In our work, we experimented using the INTERSPEECH 2024 TAUKADIAL challenge dataset consisting of Chinese and English speech samples collected while the speakers participated in picture description tasks conducted as part of a cognitive assessment protocol[142]. In the English speech recordings, the participants completed the discourse protocol and cognitive-linguistic battery with a facilitator. The discourse protocol tasks included three picture-description tasks: a) the “Cookie Theft” picture (Figure 7.1)[154], b) the “Cat Rescue” picture (Figure 7.2) [155], and c) the Norman Rockwell print “Coming and Going” (Figure 7.3) [156].

In the Chinese corpus, participants described a set of three pictures depicting Taiwanese culture. For both languages, the participants with MCI were diagnosed by experts in neuropsychology, according to the National Institute on Aging-Alzheimer's Association (NIA-AA) [157]. To avoid modeling bias, age, and gender balance were maintained across the dataset. The dataset was divided into training and test sets consisting of samples from participants in both languages, including 387 and 120 samples respectively.

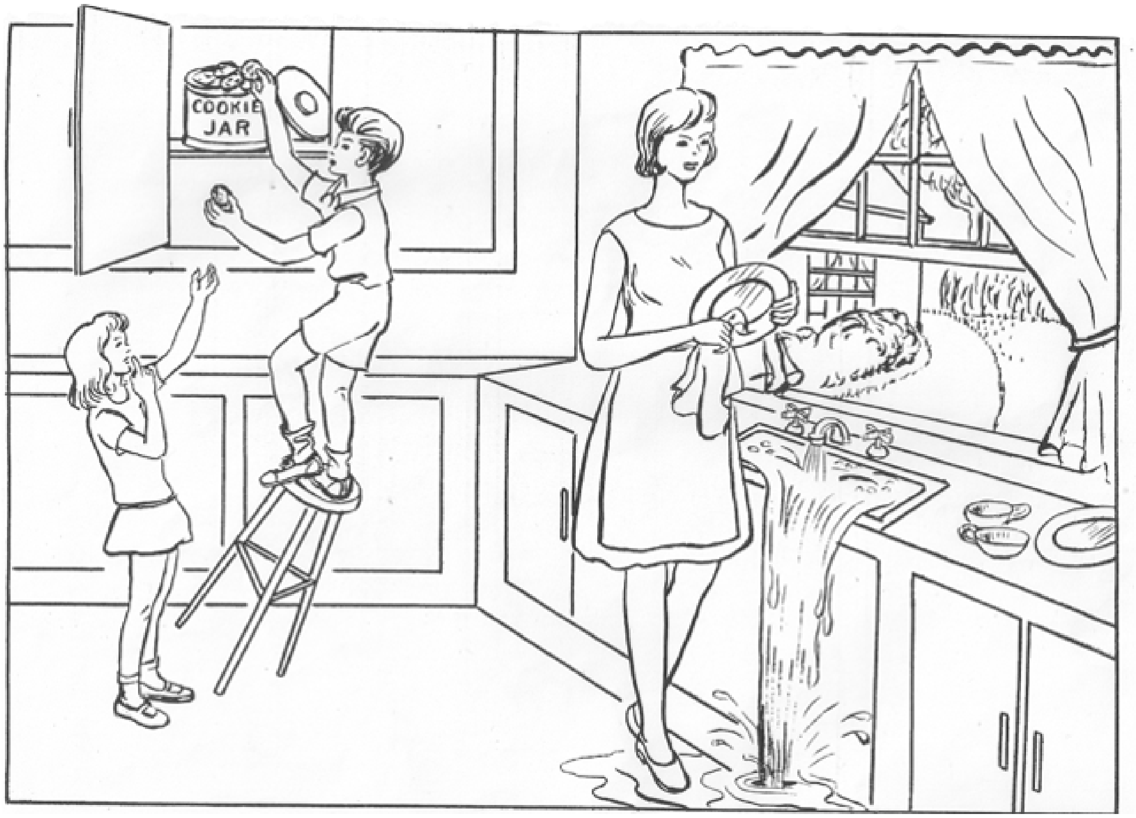


Figure 7.1: C̀ookie Theft' picture

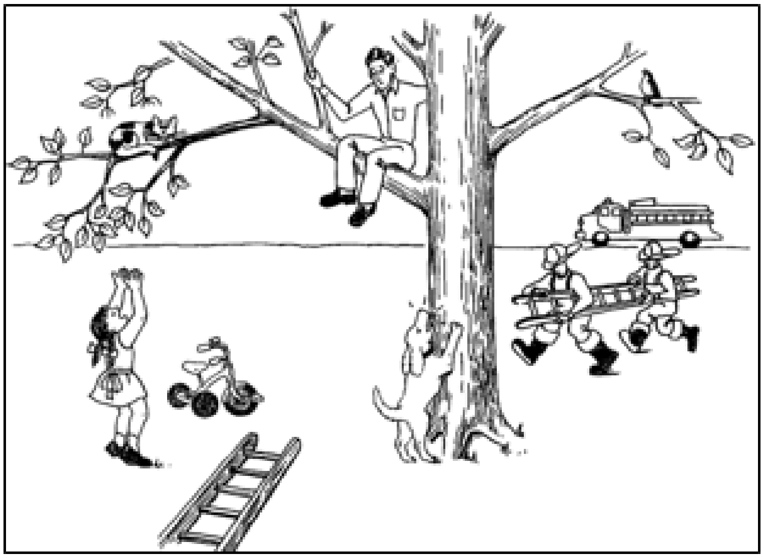


Figure 7.2: C̀at Rescue' picture

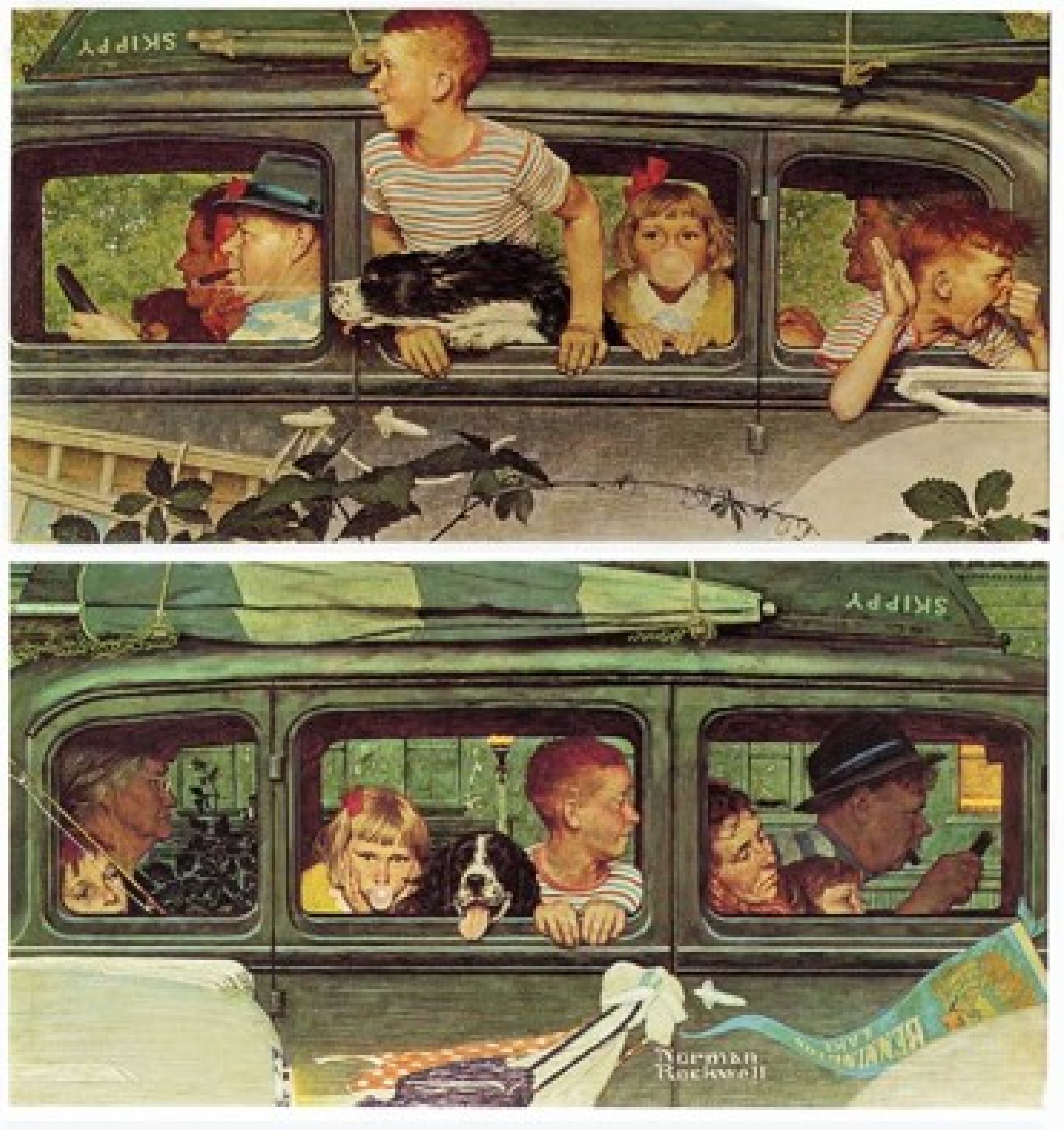


Figure 7.3: The Norman Rockwell print “Coming and Going”

7.3.2 Data Preprocessing

Language identification: The process of language identification entails categorizing audio content as either Chinese or English, constituting a binary classification task. We utilize Whisper's [158] predicted language probabilities to assign language labels to the audio. This setup guarantees that the predicted language label will exclusively be either Chinese or English for each speech sample.

Segmentation: Each audio content comprises two speakers: an interviewer and a participant. Extracting features without determining the speaker may result in weaker feature representation. Therefore, we employ the Pyannote model [159, 160] to perform speaker diarization and segmentation. The Pyannote model is trained to diarize speakers even when their audio tracks are overlapping. To leverage this capability, we first perform speaker diarization on each audio using Pyannote, assuming at most two speakers. We then extract audio segments from the original audio using Pyannote-labeled timestamps. The interviewer and participant are determined based on the length of their audio segments, assuming that the participant should speak longer than the interviewer. Each audio segment is then transcribed into text using a Whisper [158] model and previously detected language type.

7.3.3 Feature Extraction

Disfluency Features: Disfluency features in speech refer to irregularities or disruptions in the flow of speech, for example, pauses, and repetitions. These features can be indicative of MCI because they reflect underlying difficulties in language processing and memory retrieval, which are often early signs of cognitive decline. To utilize these crucial features, we developed an automatic audio segmentation method applicable to both English and Chinese samples.

The speaker diarization step results in the segmentation of the original audio sample into a sequence of P, I, and S segments, denoting the commencement and conclusion times of segments belonging to the participant, the interviewer, and periods of silence, correspondingly. Subsequently, utilizing these three distinct sequences, we compute six disfluency statistics for each sequence. These statistics include (i) the total length of each sequence, (ii) the proportion of each sequence relative to the entire duration of the audio, (iii) the average duration of sequences, (iv) the duration of the longest sequence, (v) the count of sequences, and (vi) the standard deviation of sequence lengths. Therefore, a total of 3×6 disfluency features, alongside the ratio of participant's speaking time to silent periods and the overall length of the audio recording, are employed as the feature of disfluency.

Pre-Trained Acoustic Embeddings: Pre-trained speech embeddings are representations acquired from audio data using methods tailored for sound signals. These embeddings encapsulate valuable insights about the audio, encompassing speech patterns, linguistic traits, and acoustic attributes. In our investigation, we explore a collection of audio embeddings derived from two pre-trained deep-learning models: VGGish [161] and Wav2Vec2 [162]. We examine three variations of the Wav2Vec2 model: the original wav2vec2-large-xlsr-53 model and two models fine-tuned on Chinese and English datasets exclusively. One of the fine-tuned models incorporates an all-age dataset, while the other focuses solely on elderly speakers, aligning with the demographics of our task dataset.

Pre-Trained Semantic Embeddings: Semantic features are the fundamental building blocks of meaning in language. The variations in the semantic content of speech, such as the choice of words or the complexity of expressions, can be early indicators of cognitive changes associated with MCI. We use OpenAI's `text-embedding-3-large` model [1], which has 3,072-dimensional vector space, to capture a wide array of semantic features. However, considering the constraints posed by our dataset's size, we also employ a dynamic reduction technique to scale down the embedding dimensions to 256. This approach helps us maintain the integrity of the conceptual information while mitigating the risk of overfitting for detecting MCI.

Bag-of-Audio-Words Features: We distinguish the audio segments containing participants' speech only, excluding the segments with interviewers' speech (Section 7.3.2). From the participants' speech, we extract 130 low-level descriptors (LLDs) of ComParE 2016 feature set [163], including energy, spectral, cepstral (MFCC) and voicing related LLDs, logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness from each 10 ms window using openSMILE 2.3 [164]. Bag-of-Audio-Words (BoAW) features of codebook size 500 were computed on the LLDs extracted from each speech sample using the openXBOW toolkit [165]. This feature set represents the frequency of certain acoustic patterns within each speech sample.

GPT-4 features: We utilize the GPT-4 API [166] in a zero-shot learning framework to analyze the transcripts from audio recordings. The GPT-4 model is used to examine each transcript and assign it a score ranging from 1 to 10 for a specific set of linguistic features, with 1 representing minimal presence and 10 indicating a pronounced presence of the feature within the transcript. These features are selected for their significance in cognitive and linguistic evaluations and include repetitiveness, limited vocabulary, off-topic transitions, semantic paraphasias, simplified grammar, sentence fragmentation, difficulty with naming, circumlocution, impaired comprehension, narrative coherence, and topic maintenance, resulting in the derivation of 11 distinct GPT-4 features for each audio transcript. Leveraging GPT-4's advanced capabilities in understanding natural language, we aim to capture a more detailed portrait of the participants' linguistic abilities, facilitating a comprehensive assessment of potential cognitive declines or linguistic impairments.

7.3.4 Modeling Approach

The challenge organizers specified two machine learning tasks: (a) binary classification to distinguish MCI patients from healthy controls, and (b) regression to estimate each

[1] https://platform.openai.com/docs/guides/embeddings

subject's Mini-Mental State Examination (MMSE) score. To train and validate the models, we formulate 5 cross-validation folds, maintaining the same ratio of patients and controls as the training set. We also ensure that all three samples from each individual are included in either the training or validation set.

Unimodal models

Given our different sets of features above, we explore various dimension-reduction techniques and base-learning algorithms to identify the best-performing pipeline. For dimension reduction, we apply Principal Component Analysis (PCA), Latent Semantic Analysis (LSA), and minimum redundancy maximum relevance (mRMR) [167]. The base learning models are formulated as follows.

Base Classifiers

Semantic classifier: Logistic regression (LR) model trained on GPT4 features

Fluency classifier: Neural network (NN) model trained on a combination of disfluency features and GPT4 features

Acoustic classifier #1: Random forest (RF) model trained on wav2vec features.

Acoustic classifier #2: RF model trained on VGGish features

Base Regressors:

Semantic regressor: Linear regressor trained on GPT4 features following PCA dimension reduction

Fluency regressor: XGBoost model trained on a combination of acoustic disfluency features, word disfluency features, and GPT4 features

Acoustic regressor #1: SVM model trained on VGGish features

Acoustic regressor #2: RF model trained on VGGish features

The choice of algorithms for the base unimodal models was inspired by the literature described in Section 7.2. Internal 5-fold CV was used to tune the hyperparameters for each model based on accuracy and root-mean-square-error (RMSE) for classification and regression tasks respectively.

The hyperparameters we tested were:

Dimension reduction: PCA using {10, 20, 30, 50} components, LSA using {100, 200, 500} components, and mRMR using {100, 150, 200, 250} components.

Models: (i) SVM (l2 regularization parameter: {0.1, 1, 10, 100, 1,000}, kernel: {linear, RBF, polynomial}); (ii) LR (regularization parameter: 20 values spaced evenly on a log scale in the range $[10^{-4}, 10^{4}]$, regularization type: {L1, L2}); (iii) RF (number of trees: {100, 300, 500, 700}, maximum features at each split: {5, 15, 25, 35, 45, 55}, minimum samples at leaf node: {1, 2, 3, 4}); (iv) XGBoost (maximum depth: {5, 6, 7, 8}, learning rate: {0.02, 0.05, 0.07, 0.1}, number of trees: {50, 100, 200, 500, 1,000}); and (v) NN (hidden neurons: {(20,), (20, 10), (30, 20, 10)}, initial learning rate: {0.01, 0.001, 0.0001}, and whether using early stopping).

Ensemble models

For ensembling, we adopted two approaches: *early fusion of features*, and *late-fusion of predictions of the base models*.

Early Fusion of Features: In this approach, we horizontally concatenated the disfluency features, BoAW features, demographic attributes of the participant, i.e., age and sex, and language identification (English or Chinese) for each speech sample. Collectively this feature set contains information on the participants' acoustic fluency, disfluency and demographic information relevant to their cognitive health. This step resulted in a feature vector of 522 features. Then we applied minimum-redundancy-maximum-relevance (mRMR) feature selection to select the most relevant 100, 150, 200, and 250 features for the respective task labels. For the classification task, the model exploiting 100 features yielded the highest average 5-fold CV accuracy, while for the regression we obtained the lowest root-mean-square-error (RMSE) using 250 features.

We trained a support vector machine model for the classification task (*Ensemble Classifier #1*) and a random forest model (*Ensemble Regressor #1*) for the regression task on the concatenated features. The model hyperparameters were tuned using an inner 5-fold CV within each fold. The hyperparameter configuration obtaining the highest accuracy among the folds was used to train the model on the entire training set to make predictions on the held-out test set.

Late Fusion of Models Late fusion, as an alternative method of our predictive modeling framework, aims to enhance prediction robustness and accuracy by aggregating predictions from multiple sources of information (pieces of audio) for each participant. Specifically, for each participant, we first compute individual predictions from each of the three audio samples they provided using the base models described in Section 7.3.4. These predictions consist of probabilities for classification tasks and scalar values for regression tasks. The rationale behind this strategy is to mitigate the variability and potential biases present in single audio samples, thereby harnessing a more holistic and representative insight into the participant's cognitive status.

To accomplish this, we calculate the mean of the results predicted by the base models on each of the three audio files for each participant. This entails averaging the probabilities in the context of classification to yield an averaged probability score (and then transformed to a binary prediction), and similarly, averaging the regression values to derive a single regression score. The fused binary prediction and regression score are then used as the final predictions for each participant.

To leverage the strengths of the individual base models, we ensemble the semantic, acoustic, and fluency models' predictions by averaging the classification prediction probabilities and regression values to formulate the following ensemble models:

> *Ensemble classifier #2:* Combination of *semantic classifier*, *fluency classifier* and *acoustic classifier #1*
>
> *Ensemble classifier #3:* Combination of *semantic classifier*, *fluency classifier* and *acoustic classifier #2*
>
> *Ensemble regressor #2:* Combination of *semantic regressor*, *fluency classifier* and *acoustic regressor #1*
>
> *Ensemble regressor #3:* Combination of *semantic regressor*, *fluency classifier* and *acoustic regressor #2*

Among the unimodal and ensemble models described in this section, we selected the five classifiers and regressors performing best on the 5-fold CV of the training set, to participate in the TAUKADIAL 2024 challenge. Section 7.4.2 summarizes the performance of these models on the 5-fold CV as well as on the held-out test set.

7.4 Result and Discussion

7.4.1 Performance Metrics

The MCI classification models are evaluated using the following metrics, specified by the challenge organizers:

$$\text{sensitivity}(\rho) = \frac{T_p}{T_p + F_N}$$

$$\text{specificity}(\sigma) = \frac{T_N}{T_N + F_P}$$

$$F_1 \text{ score} = \frac{2\pi\rho}{\pi + \rho}$$

where,

$$\pi = \frac{T_P}{T_P + F_P}$$

and

$$\text{Unweighted average recall (UAR)} = \frac{\sigma + \rho}{2}$$

Table 7.3: Classification performance of the models. The scores are averaged across 5-folds (20-folds for the baseline). The Rightmost column represents UAR score on the test set. The best UAR scores are marked in bold

Model	**Sensitivity**	**Specificity**	F_1	**UAR (5-fold)**	**UAR (test)**
Semantic classifier	0.784	0.618	0.758	0.702	0.568
Fluency classifier	0.821	0.546	0.759	0.683	**0.570**
Ensemble Classifier #1	0.685	0.558	0.621	0.622	0.515
Ensemble Classifier #2	0.891	0.618	0.820	**0.754**	0.563
Ensemble Classifier #3	0.891	0.600	0.815	0.745	0.515
Baseline [142]	-	-	-	0.509	0.592

Here N is the number of patients, T_P is the number of true positives, T_N is the number of true negatives, F_P is the number of false positives and F_N the number of false negatives.

The MMSE regression performance was assessed using RMSE calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

where $\hat{y_i}$ is the predicted MMSE score, and y_i is the patient's actual MMSE score.

7.4.2 Experimental Results

The performance of our proposed classifier models has been summarized in Table 7.3. On the training set, the highest classification UAR of 0.754 was achieved by our *ensemble classifier #2*, which outperforms the baseline classifiers [142]. However, on the test set, the *Fluency classifier* achieved higher UAR than the ensemble models and outperformed all but one baseline models.

On the regression task, our unimodal and ensemble regressors outperform the baseline acoustic and linguistic models on the training set. On the test set, our *Ensemble regressor #1* achieves RMSE of 2.54, which outperforms the baseline models.

7.5 Conclusion

In this work, we explore a variety of acoustic, semantic, and fluency-related features and ensemble techniques to overcome the challenge of distinguishing MCI patients from controls, alongside measuring the disorder severity on MMSE scale. Our ensemble classifier exploiting the late-fusion approach yields achieving 75.4% UAR in 5-fold CV

Table 7.4: Regression RMSE of the models. The lowest RMSE scores are marked in bold

Model	RMSE (5-folds)	RMSE (test)
Semantic regressor	2.72	3.21
Fluency regressor	2.74	3.69
Ensemble regressor#1	2.66	**2.54**
Ensemble regressor#2	**2.56**	3.23
Ensemble regressor#3	2.62	3.26
Baseline [142]	2.86	2.89

which is significantly higher than the baseline UAR and performs competitively on the test set, although the fluency model performed marginally better than the ensemble classifier on the test partition. On the other hand, the early-fusion approach performed consistently on both 5-fold CV and test sets in the regression task, outperforming the baseline models. In this approach, we provide the model with information on acoustic fluency and disfluency, coupled with relevant demographic and language identification, paving the way for consistent regression performance.

Chapter 8

Conclusion

This thesis embarks on a comprehensive exploration of speech-processing pipelines for the prediction and severity assessment of psychiatric and cognitive disorders. Through a series of six progressive experiments, significant insights are gained into the effectiveness of various preprocessing methodologies, feature sets, and machine learning algorithms.

8.1 Contributions

Throughout this thesis, six experiments have been conducted to address the following research questions essential for formulating a robust speech-processing pipeline capable of accurately predicting mental health conditions and cognitive decline:

Which aspects of voice provide the most information to infer mental health conditions?

- Analysis presented in Chapters 2 and 3 shows that depression classification accuracy was higher using comparatively **longer length spontaneous speech samples** of AVEC 2017 and AVEC 2019 datasets than shorter length guided speech samples of AVEC 2013 dataset. However, lower normalized RMSE was obtained on the AVEC 2013 dataset than the AVEC 2017 and AVEC 2019 datasets, irrespective of using feature value binning (Chapter 3) and averaging predicted scores on segments of the speech samples (Chapters 2 and 3)

- Analysis presented in Chapter 4 demonstrate that the **content** of speech samples does not show a significant impact as long as the length of speech samples is reasonably short, for example, less than one minute in the case of DEPAC dataset

- Historically, women are diagnosed with depression twice as often as men [168]. Therefore, in the experiments presented in this thesis, sex has been included an important feature to strengthen the models' predictions (Chapters 3, 4, 5).

- **Depression severity** of the subject show a weak positive correlation with the prediction error (Chapter 4)

The analysis conducted in this thesis suggests that longer, spontaneous speech samples are better for accurately identifying depression compared to shorter, guided samples. Shorter speech samples show little impact from content. Participants' sex plays a significant role in prediction accuracy, considering that women are diagnosed with depression more frequently and demonstrate symptoms in higher intensity. Additionally, there is a noticeable increase in prediction error with more severe depression symptoms. This phenomenon highlights a critical challenge in the application of such models, underscoring the necessity for additional experimentation and refinement to ensure reliable and consistent diagnostic outcomes across varying levels of symptomatology.

What are the relative merits and shortcomings of different features in predicting mental disorders and cognitive impairment?

The answer to this research question varies based on the nature of the disorder, therefore the outcomes of the analysis are summarized as follows:

Psychiatric disorders:

- **AVEC 2013 feature set** demonstrates effectiveness in predicting the prevalence and severity of depression over other popular sets of conventional acoustic features including eGeMAPS, ComParE, and emo_large (Chapter 2 and 3). Although on AVEC 2019 dataset, ComParE feature set offers a marginal improvement in regression performance from AVEC 2013 feature set, the improvement is out-weighted by the additional computational cost of constructing the ComParE feature set (6,373 features) in comparison to AVEC 2013 feature set.
- Manually curated **conventional acoustic feature** based approach achieves equal or better performance in predicting depression severity, at a remarkably less computation time and resources (Chapter 4).

Cognitive impairment:

- Combination of relevant demographics, speech disfluency, and acoustic fluency represented as BoAW features have been found effective in predicting the severity of MCI in Chapter 7.
- **Linguistic features** were identified as more informative in measuring cognitive impairment than acoustic features, as described in Chapter 6.

Manually curated conventional acoustic features prove to be equally effective or even better in predicting depression severity than deep representation features, requiring far less computational time and resources. The **AVEC 2013 feature set**, which encompasses the minimalistic eGeMAPS features and is a subset of the larger ComParE and emo_large feature sets, offers a comprehensive array of experimentally proven acoustic features tailored to capture emotional nuances in speech. Its efficacy in predicting depression prevalence and severity surpasses that of other conventional acoustic feature sets. Notably, it strikes an optimal balance between computational efficiency and predictive model performance, making it a promising tool for mental health diagnosis. While the ComParE feature set demonstrates a marginal enhancement in performance on the AVEC 2019 dataset, its adoption entails significantly higher computational overhead.

From the analysis on predicting cognitive disorders, a combination of demographics, speech disfluency, and acoustic fluency features is found effective in predicting the severity of MCI, with linguistic features being identified as more informative in measuring cognitive impairment compared to acoustic features.

Which algorithms perform best in inferring indicators of mental and neurocognitive disorders?

This thesis investigates a wide range of probabilistic, non-linear, and ensemble machine learning algorithms to discern their effectiveness in inferring indicators of mental and neurocognitive disorders (Table 1.3). Among these algorithms, basic models such as random forest and hidden Markov model exhibit strong performance in classifying depressed individuals from healthy ones and predicting depression severity scores (Chapters 2, 3). Random forest has been established as the most effective regressor to predict neurocognitive impairment (Chapters 6, 7) as well. Additionally, feedforward neural networks prove adept at forecasting depression severity (Chapter 4). In recent experiments, 1-dimensional convolutional neural networks show promise in enhancing prediction accuracy (Chapter 5). However, further exploration into more complex time series models tailored for personalized predictions remains an avenue for further investigation.

8.2 Implications and Future Directions

The findings presented in this thesis hold significant implications for both research and practical applications in mental and cognitive health assessment. The adaptability and performance of the speech-processing pipeline demonstrated across various datasets and disorders underscore the potential of the proposed pipeline in revolutionizing mental health diagnostics.

Moving forward, I would like to explore further ways to optimize feature selection, model architectures, and dataset curation. In particular, I envision identifying how to personalize the model to monitor psychiatric and neurodegenerative diseases. In order to do so, what should be the characteristics of the training corpus? Investigating language-agnostic systems for global mental healthcare dissemination could also yield valuable insights. Additionally, the integration of multimodal data sources and advanced machine-learning techniques presents an exciting avenue for future research. Moreover, the scalability and generalizability of the proposed pipeline should be rigorously tested across diverse populations and cultural contexts to ensure equitable access to mental health services.

In conclusion, this thesis contributes to the growing body of literature in computational psychiatry and offers a promising framework for leveraging speech data in mental health assessment. Bridging the gap between traditional diagnostic methods and cutting-edge machine learning techniques, this work paves the way for more efficient, accessible, and personalized mental health care interventions.

Bibliography

[1] *Depressive disorder*, https://www.who.int/news-room/fact-sheets/detail/depression.

[2] *Dementia*, https://www.who.int/news-room/fact-sheets/detail/dementia.

[3] J. J. McGrath *et al.*, "Age of onset and cumulative risk of mental disorders: A cross-national analysis of population surveys from 29 countries," *The Lancet Psychiatry*, vol. 10, no. 9, pp. 668–681, 2023.

[4] S. Banovic, L. J. Zunic, and O. Sinanovic, "Communication difficulties as a result of dementia," *Materia socio-medica*, vol. 30, no. 3, p. 221, 2018.

[5] G. R. Kuperberg, "Language in schizophrenia part 1: An introduction," *Language and linguistics compass*, vol. 4, no. 8, pp. 576–589, 2010.

[6] *Speech & swallowing in parkinson's*, https://www.parkinson.org/sites/default/files/documents/Speech-Swallowing-2024_1.pdf, 2024.

[7] A. Harati, E. Shriberg, T. Rutowski, P. Chlebek, Y. Lu, and R. Oliveira, "Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 7273–7277.

[8] J. G. Borkowski, A. L. Benton, and O. Spreen, "Word fluency and brain damage," *Neuropsychologia*, vol. 5, no. 2, pp. 135–140, 1967.

[9] H. Jiang *et al.*, "Investigation of different speech types and emotions for detecting depression using different classifiers," *Speech Communication*, vol. 90, pp. 39–46, 2017.

[10] P. Fossati, A.-M. Ergis, J.-F. Allilaire, *et al.*, "Qualitative analysis of verbal fluency in depression," *Psychiatry research*, vol. 117, no. 1, pp. 17–24, 2003.

[11] R. N. Trifu, B. NEMEŞ, C. Bodea-Hațegan, and D. Cozman, "Linguistic indicators of language in major depressive disorder (mdd). an evidence based research.," *Journal of Evidence-Based Psychotherapies*, vol. 17, no. 1, 2017.

[12] *Depression*, http://www.who.int/mediacentre/factsheets/fs369/en/.

[13] R. Wang *et al.*, "StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, 2014, pp. 3–14.

[14] M. Valstar *et al.*, "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.

[15] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "A study of acoustic features for the classification of depressed speech," in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, IEEE, 2014, pp. 1331–1335.

[16] F. Ringeval *et al.*, "AVEC 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ACM, 2017, pp. 3–9.

[17] 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, 2008.

[18] M. H. Sanchez *et al.*, "Using prosodic and spectral features in detecting depression in elderly males," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[19] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2011.

[20] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling spectral variability for the classification of depressed speech.," in *Interspeech*, 2013, pp. 857–861.

[21] K. C. Fraser, F. Rudzicz, and G. Hirst, "Detecting late-life depression in alzheimer's disease through analysis of speech and language," in *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, 2016, pp. 1–11.

[22] 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.

[23] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of biomedical informatics*, 2018.

[24] Y. Özkanca, C. Demiroglu, A. Besirli, and S. Celik, "Multi-lingual depression-level assessment from conversational speech using acoustic and text features," *Proc. Interspeech 2018*, pp. 3398–3402, 2018.

[25] M. R. Morales, "Multimodal depression detection: An investigation of features and fusion techniques for automated systems," 2018.

[26] B. Sun *et al.*, "A random forest regression method with selected-text feature for depression assessment," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ACM, 2017, pp. 61–68.

[27] S. Dham, A. Sharma, and A. Dhall, "Depression scale recognition from audio, visual and text analysis," *arXiv preprint arXiv:1709.05865*, 2017.

[28] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ACM, 2017, pp. 45–51.

[29] A. Samareh, Y. Jin, Z. Wang, X. Chang, and S. Huang, "Predicting depression severity by multi-modal feature engineering and fusion," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[30] 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.

[31] *Depression*, http://www.apa.org/topics/depression.

[32] *Suicide data*, https://www.who.int/news-room/fact-sheets/detail/suicide.

[33] *Depression, a hidden burden*, http://www.who.int/mental_health/management/ \protect\@normalcr\relaxdepression/flyer_depression_2012.pdf.

[34] 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.

[35] J. K. Darby, N. Simmons, and P. A. Berger, "Speech and voice parameters of depression: A pilot study," *Journal of Communication Disorders*, vol. 17, no. 2, pp. 75–85, 1984.

[36] 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.

[37] H.-M. Liu, F.-M. Tsao, and P. K. Kuhl, "The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy," *The Journal of the Acoustical Society of America*, vol. 117, no. 6, pp. 3879–3889, 2005.

[38] G. S. Turner, K. Tjaden, and G. Weismer, "The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 5, pp. 1001–1013, 1995.

[39] P. A. McRae, K. Tjaden, and B. Schoonings, "Acoustic and perceptual consequences of articulatory rate change in Parkinson disease," *Journal of Speech, Language, and Hearing Research*, 2002.

[40] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews.," in *LREC*, Citeseer, 2014, pp. 3123–3128.

[41] D. DeVault *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 International Conference on Autonomous agents and multi-agent systems*, 2014, pp. 1061–1068.

[42] T. Laosaphan and T. Yingthawornsuk, *Classification of depressed speakers based on MFCC in speech samples*, 2012.

[43] 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.

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

[45] 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.

[46] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[47] A. Afshan, J. Guo, S. J. Park, V. Ravi, J. Flint, and A. Alwan, "Effectiveness of voice quality features in detecting depression.," in *Interspeech*, 2018, pp. 1676–1680.

[48] A. Salekin, J. W. Eberle, J. J. Glenn, B. A. Teachman, and J. A. Stankovic, "A weakly supervised learning framework for detecting social anxiety and depression," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 2, pp. 1–26, 2018.

[49] M. Muzammel, H. Salam, Y. Hoffmann, M. Chetouani, and A. Othmani, "Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis," *Machine Learning with Applications*, vol. 2, p. 100 005, 2020.

[50] N. Seneviratne, J. R. Williamson, A. C. Lammert, T. F. Quatieri, and C. Espy-Wilson, "Extended study on the use of vocal tract variables to quantify neuromotor coordination in depression," *Proc. Interspeech 2020*, pp. 4551–4555, 2020.

[51] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 668–680, 2017.

[52] Z. Zhao *et al.*, "Hybrid network feature extraction for depression assessment from speech," *Proc. Interspeech 2020*, pp. 4956–4960, 2020.

[53] J. R. Williamson *et al.*, "Detecting depression using vocal, facial and semantic communication cues," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 11–18.

[54] 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.

[55] A. Haque, M. Guo, A. S. Miner, and L. Fei-Fei, "Measuring depression symptom severity from spoken language and 3d facial expressions," *arXiv preprint arXiv:1811.08592*, 2018.

[56] L. Yang, D. Jiang, and H. Sahli, "Feature augmenting networks for improving depression severity estimation from speech signals," *IEEE Access*, vol. 8, pp. 24 033–24 045, 2020.

[57] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression severity prediction based on biomarkers of psychomotor retardation," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 37–43.

[58] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 81–88.

[59] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 55–63.

[60] L. Zhang, R. Duvvuri, K. K. Chandra, T. Nguyen, and R. H. Ghomi, "Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative," *Depression and anxiety*, 2020.

[61] W. Fan, Z. He, X. Xing, B. Cai, and W. Lu, "Multi-modality depression detection via multi-scale temporal dilated cnns," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 73–80.

[62] F. Ringeval *et al.*, "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 3–12.

[63] S. P. Dubagunta, B. Vlasenko, and M. M. Doss, "Learning voice source related information for depression detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6525–6529.

[64] 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.

[65] 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.

[66] S. Yin, C. Liang, H. Ding, and S. Wang, "A multi-modal hierarchical recurrent neural network for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 65–71.

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

[68] S. Scherer, G. M. Lucas, J. Gratch, A. S. Rizzo, and L.-P. Morency, "Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 59–73, 2016.

[69] B. Schuller *et al.*, "The interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[70] B. Schuller *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016), VOLS 1-5*, 2016, pp. 2001–2005.

[71] M. Tasnim and E. Stroulia, "Detecting depression from voice," in *Canadian Conference on Artificial Intelligence*, Springer, 2019, pp. 472–478.

[72] B. Schuller *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.

[73] F. Eyben, *Real-time speech and music classification by large audio feature space extraction.* Springer, 2015.

[74] F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.

[75] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[76] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[77] A. Cuevas, M. Febrero, and R. Fraiman, "An anova test for functional data," *Computational statistics & data analysis*, vol. 47, no. 1, pp. 111–122, 2004.

[78] M. Pal, "Random forest classifier for remote sensing classification," *International journal of remote sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[79] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[80] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, "Xgboost: Extreme gradient boosting," *R package version 0.4-2*, pp. 1–4, 2015.

[81] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[82] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp. 275–285, 2004.

[83] L. Rabiner and B. Juang, "An introduction to hidden markov models," *ieee assp magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[84] K. Gurney, *An introduction to neural networks.* CRC press, 1997.

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

[86] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[87] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[88] C. M. Bishop, *Pattern recognition and machine learning.* Springer, 2006.

[89] M. Tasnim, M. Ehghaghi, B. Diep, and J. Novikova, "Depac: A corpus for depression and anxiety detection from speech," in *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, 2022, pp. 1–16.

[90] 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*, 2016, pp. 35–42.

[91] B. Diep, M. Stanojevic, and J. Novikova, *Multi-modal deep learning system for depression and anxiety detection*, 2022. arXiv: 2212.14490 [cs.SD].

[92] B. Yalamanchili, N. S. Kota, M. S. Abbaraju, V. S. S. Nadella, and S. V. Alluri, "Real-time acoustic based depression detection using machine learning techniques," in *2020 International conference on emerging trends in information technology and engineering (ic-ETITE)*, IEEE, 2020, pp. 1–6.

[93] A. Balagopalan and J. Novikova, "Comparing Acoustic-Based Approaches for Alzheimer's Disease Detection," in *Proc. Interspeech 2021*, 2021, pp. 3800–3804. DOI: 10.21437/Interspeech.2021-759.

[94] M. Valstar *et al.*, "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.

[95] 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.

[96] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[97] J. Novikova and K. Shkaruta, *DECK: Behavioral Tests to Improve Interpretability and Generalizability of BERT Models Detecting Depression from Text*, 2022. arXiv: 2209.05286 `[cs.CL]`.

[98] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Hierarchical attention transfer networks for depression assessment from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7159–7163.

[99] L. Yang, D. Jiang, W. Han, and H. Sahli, "Dcnn and dnn based multi-modal depression recognition," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2017, pp. 484–489.

[100] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[101] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, IEEE, vol. 1, 2006, pp. I–I.

[102] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[103] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[104] S. Amiriparian *et al.*, "Snore sound classification using image-based deep spectrum features," en, in *Interspeech 2017*, ISCA, Aug. 2017, pp. 3512–3516.

[105] Z. Zhao, R. Anand, and M. Wang, "Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform," in *2019 IEEE international conference on data science and advanced analytics (DSAA)*, IEEE, 2019, pp. 442–452.

[106] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The adresso challenge," *arXiv preprint arXiv:2104.09356*, 2021.

[107] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12 592–12 594, 2020.

[108] M. Kappen, M.-A. Vanderhasselt, and G. M. Slavich, "Speech as a promising biosignal in precision psychiatry," *Neuroscience & Biobehavioral Reviews*, p. 105 121, 2023.

[109] Z. Shah *et al.*, "Exploring Language-Agnostic Speech Representations Using Domain Knowledge for Detecting Alzheimer's Dementia," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–2. DOI: 10.1109/ICASSP49357.2023.10095593.

[110] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[111] A. Othmani, D. Kadoch, K. Bentounes, E. Rejaibi, R. Alfred, and A. Hadid, "Towards robust deep neural networks for affect and depression recognition from speech," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, Springer, 2021, pp. 5–19.

[112] B. Diep, M. Stanojevic, and J. Novikova, "Multi-modal deep learning system for depression and anxiety detection," in *Empowering Communities: A Participatory Approach to AI for Mental Health*, 2022.

[113] D. F. Santomauro *et al.*, "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic," *The Lancet*, vol. 398, no. 10312, pp. 1700–1712, 2021.

[114] S. Weinberger, "Speech accent archive," *George Mason University*, 2015.

[115] S. H. Lovibond, "Manual for the depression anxiety stress scales," *Sydney psychology foundation*, 1995.

[116] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, e1008228, 2020.

[117] A. Y. Kim, E. H. Jang, S.-H. Lee, K.-Y. Choi, J. G. Park, and H.-C. Shin, "Automatic Depression Detection Using Smartphone-Based Text-Dependent Speech Signals: Deep Convolutional Neural Network Approach," *Journal of Medical Internet Research*, vol. 25, e34474, 2023.

[118] M. Tasnim and J. Novikova, "Cost-effective models for detecting depression from speech," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2022, pp. 1687–1694.

[119] A. Fatima, Y. Li, T. T. Hills, and M. Stella, "DASentimental: Detecting depression, anxiety, and stress in texts via emotional recall, cognitive networks, and machine learning," *Big Data and Cognitive Computing*, vol. 5, no. 4, p. 77, 2021.

[120] A. Association, "2019 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 15, no. 3, pp. 321–387, 2019.

[121] K. Ritchie and S. Lovestone, "The dementias," *The Lancet*, vol. 360, no. 9347, pp. 1759–1766, 2002.

[122] H. Braak and E. Braak, "Neuropathological stageing of alzheimer-related changes," *Acta neuropathologica*, vol. 82, no. 4, pp. 239–259, 1991.

[123] R. D. Terry *et al.*, "Physical basis of cognitive alterations in alzheimer's disease: Synapse loss is the major correlate of cognitive impairment," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 30, no. 4, pp. 572–580, 1991.

[124] P. J. Nestor, P. Scheltens, and J. R. Hodges, "Advances in the early detection of alzheimer's disease," *Nature medicine*, vol. 10, no. 7, S34–S41, 2004.

[125] J. Weller and A. Budson, "Current understanding of alzheimer's disease diagnosis and treatment," *F1000Research*, vol. 7, 2018.

[126] E. Arnáiz and O. Almkvist, "Neuropsychological features of mild cognitive impairment and preclinical alzheimer's disease," *Acta Neurologica Scandinavica*, vol. 107, pp. 34–41, 2003.

[127] D. M. Jacobs, M. Sano, G. Dooneief, K. Marder, K. L. Bell, and Y. Stern, "Neuropsychological detection and characterization of preclinical alzheimer's disease," *Neurology*, vol. 45, no. 5, pp. 957–962, 1995.

[128] Y.-W. Chien, S.-Y. Hong, W.-T. Cheah, L.-H. Yao, Y.-L. Chang, and L.-C. Fu, "An automatic assessment system for alzheimer's disease based on speech using feature sequence generator and recurrent neural network," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[129] I. Hoffmann, D. Nemeth, C. D. Dye, M. Pákáski, T. Irinyi, and J. Kálmán, "Temporal parameters of spontaneous speech in alzheimer's disease," *International journal of speech-language pathology*, vol. 12, no. 1, pp. 29–34, 2010.

[130] 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.

[131] L. Tóth *et al.*, "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.

[132] S. Mirzaei *et al.*, "Automatic speech analysis for early Alzheimer's disease diagnosis," in *JETSAN 2017 : 6e Journées d'Etudes sur la Télésanté*, Bourges, France, May 2017, pp. 114 –116. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01618834.

[133] S. Wankerl, E. Nöth, and S. Evert, "An n-gram based approach to the automatic diagnosis of alzheimer's disease from spoken language.," in *INTERSPEECH*, 2017, pp. 3162–3166.

[134] D. Kempler, "Language changes in dementia of the alzheimer type," *Dementia and communication*, pp. 98–114, 1995.

[135] H. S. Kirshner, "Primary progressive aphasia and alzheimer's disease: Brief history, recent evidence," *Current neurology and neuroscience reports*, vol. 12, no. 6, pp. 709–714, 2012.

[136] A.-L. R. Adlam, S. Bozeat, R. Arnold, P. Watson, and J. R. Hodges, "Semantic knowledge in mild cognitive impairment and mild alzheimer's disease," *Cortex*, vol. 42, no. 5, pp. 675–684, 2006.

[137] M. Nicholas, L. K. Obler, M. L. Albert, and N. Helm-Estabrooks, "Empty speech in alzheimer's disease and fluent aphasia," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 3, pp. 405–410, 1985.

[138] 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.

[139] B. MacWhinney, *Tools for analyzing talk part 1: The chat transcription format.*

[140] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[141] R. Chakraborty, M. Pandharipande, C. Bhat, and S. K. Kopparapu, "Identification of dementia using audio biomarkers," *arXiv preprint arXiv:2002.12788*, 2020.

[142] S. Luz *et al.*, "Connected Speech-Based Cognitive Assessment in Chinese and English," DOI: 10.48550/ARXIV.2404.nnnnn (TBA), arXiv, 2024. DOI: 10.48550/ARXIV.2404.nnnnn.

[143] I. Vigo, L. Coelho, and S. Reis, "Speech-and language-based classification of Alzheimer's disease: a systematic review," *Bioengineering*, vol. 9, no. 1, p. 27, 2022.

[144] 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.

[145] C. Themistocleous, M. Eckerström, and D. Kokkinakis, "Identification of mild cognitive impairment from speech in Swedish using deep sequential neural networks," *Frontiers in neurology*, vol. 9, p. 412 560, 2018.

[146] T. Wang *et al.*, “Identification of mild cognitive impairment among Chinese based on multiple spoken tasks,” *Journal of Alzheimer's Disease*, vol. 82, no. 1, pp. 185–204, 2021.

[147] L. Calzà, G. Gagliardi, R. R. Favretti, and F. Tamburini, “Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia,” *Computer Speech & Language*, vol. 65, p. 101 113, 2021.

[148] V. Vincze *et al.*, “Detecting mild cognitive impairment by exploiting linguistic information from transcripts,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, pp. 181–187.

[149] 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.

[150] L. Jin *et al.*, “Consen: Complementary and simultaneous ensemble for Alzheimer's disease detection and mmse score prediction,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–2.

[151] B. Tamm, R. Vandenberghe, and H. Van Hamme, “Cross-lingual transfer learning for Alzheimer's detection from spontaneous speech,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–2.

[152] K. Mei *et al.*, “The USTC system for ADReSS-M challenge,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–2.

[153] Z. Shah *et al.*, “Exploring language-agnostic speech representations using domain knowledge for detecting Alzheimer's dementia,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–2.

[154] H. Goodglass and E. Kaplan, *Boston diagnostic aphasia examination booklet.* Lea & Febiger, 1983.

[155] L. E. Nicholas and R. H. Brookshire, “A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia,” *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 338–350, 1993.

[156] N Rockwell, “Going and coming [Oil on canvas],” *Norman Rockwell Art Collection Trust, Indianapolis, IN, United States*, 1947.

[157] M. S. Albert *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*, vol. 7, no. 3, pp. 270–279, 2011.

[158] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust Speech Recognition via Large-Scale Weak Supervision*, 2022. arXiv: 2212.04356 `[eess.AS]`.

[159] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. INTERSPEECH 2023*, 2023.

[160] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proc. INTERSPEECH 2023*, 2023.

[161] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 131–135.

[162] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[163] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, p. 292, 2013.

[164] F. Eyben, F. Weninger, M. Wöllmer, and B Shuller, "Open-source media interpretation by large feature-space extraction," *TU Munchen, MMK*, 2016.

[165] M. Schmitt and B. Schuller, "OpenxBoW: introducing the passau open-source crossmodal bag-of-words toolkit," 2017.

[166] J. Achiam *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[167] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[168] *Is depression in men overlooked?* https://cihr-irsc.gc.ca/e/48856.html.