Project Plagiarism Coversheet

This coversheet must be added to your submission to turnitin, as part of one dissertation submission.

Title of Project	Exploration of changes in prosodic and voice quality markers associated with depression severity
Supervisor	Dr Nicholas Cummins
Deadline	16th September 2022

Statement of Academic Integrity

I have read the College Regulations on Cheating and Plagiarism and state that this piece of work is my own and does not contain any unacknowledged work from any other hand-written, printed or internet source. I further state that I have acknowledged all sources used and quoted in the attached work. I understand the penalties that can be applied in cases of proven cheating and plagiarism.

I have submitted a Pdf file of the entire project, names “K20107532_ ASMHI Research Project”, **to Turnitin**	**YES**
Surname	Metelo Calhas Ferreira
First Name	Sara
Student Signature	DocuSigned by: Sara Metelo Calhas Ferreira EEB9B722BCB24F3...
Date	16th September 2022

Exploration of changes in prosodic and voice quality markers associated with depression severity

MSc Applied Statistical Modelling and Health Informatics
King's College London

16th September 2022

Module Name: ASMHI Research Project
Module Code: 7PAVREPR
Student Number: K20107532
Supervisor (s): Dr Nicholas Cummins
Word Count: 14956 (including abstract, table of contents, list of tables, list of figures, list of acronyms, all tables, and figures. Excludes references and appendices).

Abstract

Background: This project focused on exploring changes in speech prosody and voice acoustics features that could act as biomarkers in discriminating depression severity. The project specifically aimed to answer the following questions: a) what changes in prosodic and voice quality markers are associated with depression severity? b) can speech features be used as a biomarker to classify depression severity?

Methods: The project used a dataset from the Remote Assessment of Disease and Relapse-Central Nervous System (RADAR-CNS) international research consortium (Radar-CNS, 2016). A total of 585 patients in three cohorts, United Kingdom, Spain, and the Netherlands, were part of the dataset. From the total number of participants, those that provided less than three speech recordings during the longitudinal study have been excluded from the analysis. To answer the first research question, linear mixed effects models were developed to identify key speech biomarkers of depression severity per cohort. To answer the second question of the research, the following Machine Learning classifiers were built: SVMs, Random Forest, XGBoost, Feedforward Neural Networks and Convolutional Neural Networks.

Results: Specific voice features associated with severity of depression were lower speech duration, higher number of pauses, slower articulation rates, lower values of phonation ratio, lower pitch (F0) mean values, lower levels of loudness and high Jitter. These were however not significant across all the cohorts. Of distinct difference from previous findings in the literature were higher speaking rates, low pause frequency and higher HNR values. It was also concluded that speech from scripted speech tasks provided better insights to discriminate depression severity than speech from unscripted tasks. Machine learning classifiers built with speech features from scripted speech tasks together with other predictors such as demographics, baseline depression and anxiety scores performed strongly in classifying severity of depression.

Table of contents

List of Tables

List of Figures

1. Introduction

Depression is a psychiatric disorder affecting approximately one in five adults in the UK (Williams et al., 2021). Depression and anxiety are estimated to cost 1 trillion USD$ each year at a global level according to the World Health organization (2022). These psychiatric disorders have devastating consequences, with an estimated 300,000 individuals losing their job every year in the UK alone due to mental health problems (Stevenson & Farmer, 2017). With the societal challenges posed by Covid-19 in 2020, and its implications in areas as diverse as peoples' lifestyle, social interactions, work dynamics, family support structures, education, evidence suggests that mental health has worsened even further over the last two years. Latest figures from the UK Mental Health foundation (2021), suggest that suicidal ideation has risen during 2020 and 2021 from 8% in April 2020 to 12% in November 2021 (amongst a sample of more than 4000 UK adults) with the latest figures showing that c. 33% of the population in the UK felt anxious, being this figure 40% for people with pre-existing mental health problems. The existing clinical care pathways to treat patients with mental health disorders are scarce. Results from a study examining the treatment gap in mental health care conducted by Kohn et al. (2004) revealed an estimated 70% of the global population has limited access to appropriate treatment, suggesting that standard clinical pathways are not able to cope with this high demand.

As a result, diagnostic methods that offer early and objective depression diagnosis are needed now more than ever before. To date, the available diagnostics for depression tend to rely on subjective patients' self-assessments and diagnosis by clinicians involving tools such as the Hamilton Depression Rating Scale, Montgomery-Asberg Depression Rating Scale or Inventory of Depressive Symptomatology, which may sometimes lack in objectivity due to recall biases (Alghowinem et al., 2013; Cummins et al., 2015, 2020; Huang et al., 2019; Mundt et al., 2007, 2012). In addition, it is forecasted a staggering 50% of diagnosis are missed through these methods (Cummins et al., 2015).

Physiological biomarkers such as changes in body movement, sleep patterns, speech, mood can act as predictors of depression onset or relapse (Doryab et al., 2014; Jeong et al., 2016; Matcham et al., 2019). However, as individuals don't necessarily notice these changes or have difficulties in recalling them, the correct time to intervene with treatment may be missed (Matcham et al., 2019). New technologies such as smartphones, sensors, fitness trackers and voice assistants have been exponentially increasing in consumer adoption (Huang et al., 2019; Jeong et al., 2016; Steinhubl et al., 2015). This new paradigm opens an opportunity to collect real time data on these physiological biomarkers, ultimately building an individual digital phenotype which can help depression diagnosis to be objective and preventative.

2. Research objectives

Over the past years, speech has been growing as a potential biomarker to offer an objective and timely diagnosis for depression in a clinical setting (Cummins et al., 2015; Horwitz et al., 2013; Low et al., 2020; Yamamoto et al., 2020, 2020; Yang et al., 2013). Previous studies showed significant results for prosodic, voice quality and articulatory features in differentiating healthy and individuals with depression (Cummins et al., 2015; Low et al., 2020). These studies, however, have relied on small sample sizes, have been largely cross sectional in nature, mostly conducted in laboratory/controlled settings and lacked a standardisation of protocols to collect the data (Cummins et al., 2015; Huang et al., 2019; Low et al., 2020).

The new advances in technology described earlier, allowing the creation of digital phenotypes in real time, open a great opportunity to explore the use of these remote measurement technologies (Cummins et al., 2020; Huang et al., 2019; Low et al., 2020) as a non-intrusive and more objective form of understanding changes in underlying physiological biomarkers that can be used as predictors of Depression severity.

The present research work aims to further advance speech science by identifying speech features that act as biomarkers in discriminating depression severity and allow to predict severity onset of Major Depressive Disorder. To do so, the project will use a dataset from the Remote Assessment of Disease and Relapse-Central Nervous System (RADAR-CNS) international research consortium (Radar-CNS, 2016). To the best of the author's knowledge, this is the largest longitudinal cohort study done to date on depression relapse using remote assessment technologies (RMT) across different languages (English, Dutch, and Spanish) and over a long period of time (18 months in this case). This dataset offers an unprecedent opportunity to further advance the theory on depression assessment and prediction by validating previously speech biomarkers found in controlled laboratory experiments, that can be used as a proxy for depression diagnosis and monitoring.

The project research questions and hypotheses to test will be introduced in detail in section 4 of this document, after the literature review on speech science theory as a depression biomarker.

3. Literature review: Speech as a biomarker for depression

Speech is a complex process involving a wide number of cognitive mechanisms such as muscle coordination, language comprehension, language production, executive functions, visual and auditory areas (Low et al., 2020). Because of its complexity and observed changes in speech linked with depression, as part of diagnosis clinicians also use a patient spoken language and quality of speech as a method to evaluate mental health status (Stasak et al., 2019). Research suggests changes in speech have been associated with depression symptoms (Alghowinem et al., 2013; Cummins et al., 2015; Darby et al., 1984; Mundt et al., 2007, 2012; Teixeira et al., 2013).

Areas such as changes in prosody and voice quality have shown to be predictors of depression (Darby et al., 1984; Mundt et al., 2007, 2012). In prosody, depressed individuals show greater speech pause times (Alghowinem et al., 2013; Huang et al., 2019; Mundt et al., 2007, 2012; Stasak et al., 2019), lower speech/pause ratio, as well as slower speaking rates (Cannizzaro et al., 2004; Mundt et al., 2007, 2012). Mundt et al. (2007, 2012) also revealed these individuals to have longer speech recordings lengths when performing laboratory tests versus healthy individuals. This is in large believed to be due to the longer pauses.

Contrary to some studies, such as the ones from Mundt et al. (2007, 2012), in a research by Albuquerque et al. (2021) total speech duration was found to be shorter amongst depressed individuals. The authors also found depressed individuals to have higher vowel duration (phonation time), higher pause duration and a decrease in the number of syllables, which according to the authors can explain the lower total speech duration for these individuals.

In a more recent study by Yamamoto et al. (2020) amongst 241 individuals (some with Major depressive disorder, some with Bipolar disorder and healthy controls), prosody features such as speech pause times (longer for depressed individuals), slower speaking rates and longer response times were significantly associated with depressed individuals in a spontaneous speech setting (i.e., during interviews about everyday topics).

The same prosody features also reveal to have different significance levels in being depression biomarkers if derived from automated speech or from free speech. By automated, scripted or read speech, it is meant speech derived from reading a passage or for example counting from a predefined list, whilst by free or spontaneous speech it is meant the dialogues we have in the wild, i.e., dialogues between people or for example whilst expressing one's view without being reading from a predetermined material.

In a study by Mundt et al. (2007) amongst 35 patients suffering from depression it was found that total pause time during automated/read speech was more correlated with depression than total pause time during free/spontaneous speech. In contrast, in the same study it was found that speech pause variability and speech/pause ratio can be better derived from free speech.

Alghowinem et al. (2013) in their study to understand the difference between automated speech and free speech in discriminating individuals with depression, found that in general spontaneous speech voice features were significantly better in classifying depression.

However, in the case of voice quality biomarkers such as F0 and shimmer, speech from read speech was giving better results. In addition, the same study revealed the use of thin slicing for the automated speech (i.e., using just the beginning of each sentence) offered better results than analysing the entire speech.

On measures of voice quality, Teixeira et al. (2013) and Cummins et al. (2015) argue that specific measures can help in detecting pathological speech, these are: fundamental frequency (e.g., F0), jitter, shimmer and harmonic-to-noise ratio (HNR).

The fundamental frequency, F0 (given in Hz), measures the number of times a sound wave repeats over a specific time (i.e., in speech science it refers to the rate of vibration of the vocal folds during voiced speech production) and tends to be in the lower range in depressed individuals, thus indicating a monotone speech (Alghowinem et al., 2013; Low et al., 2020). However, mixed results are found in the literature for the role of F0 in discriminating depression. For example, whilst some authors (Darby et al., 1984; Mundt et al., 2007) found statistically significant evidence for the role a low F0 plays in discriminating depression, other authors (Mundt et al., 2012; Yang et al., 2013) found no evidence for its role in its association with depression. Yang et al. (2013) in their study to understand vocal prosody changes with severity of depression, advanced F0 to be more linked with personality traits rather than with depressive severity symptoms.

As per F0 results in the literature, loudness is also a voice acoustics feature with mixed results for significance association with depressive symptoms. In a study by Albuquerque et al. (2021), loudness contrary to the authors hypothesis was found to be higher in depressed individuals.

Jitter measures frequency variation from cycle to cycle and tends to be higher for depressed individuals (Alghowinem et al., 2013; Teixeira et al., 2013). Shimmer measures the amplitude of the sound wave and tends to be lower in individuals suffering from depression (Teixeira et al., 2013). HNR measures the efficiency of speech given by two components: vibration of the vocal cords and the noise from the glottis (Alghowinem et al., 2013; Teixeira et al., 2013). A low HNR tends to be associated with depression due to differences in the quality of air flow in speech production in depressed individuals (Teixeira et al., 2013).

On measures of formant features, depressive speech tends to investigate F1 and F2 (Cummins et al., 2015). Formant features in speech science refer to the frequency components of the speech signal. F1 refers to the lowest frequency (associated with vowels), whilst F2 refers to the second one (also associated with vowels). In a study by Flint et al. (1993) amongst 30 individuals suffering from major depressive disorder, the second formant feature was significantly associated with severity of depression. These individuals revealed a decrease in F2 versus healthy subjects. However, this is not consistent across other studies such as Mundt et al. (2012), where F2 is not significantly associated with depression severity.

Lastly, spectral energy is often analysed when looking into depression. As seen with F0, F1 and F2, energy levels tend to have mixed results in studies done to date (Cummins et al., 2015). Whilst some studies clearly find correlation between spectral energy and depression, others do not find a significance.

4. Research questions and hypotheses to be tested

4.1. Research questions

The main research question explored in this project is: **what changes in prosodic and voice quality markers are associated with depression severity?** In answering this question, several hypotheses will be tested with the aim to better understand which prosody and voice acoustics features are involved in depression severity.

In addition, work undertaken in this project also aims to understand if speech features can serve as a good basis to classify individuals with severe levels of depression from those with mild symptoms. The second research question is therefore: **can speech features be used as a biomarker to classify depression severity?**

4.2. Hypotheses to be tested

The hypotheses to test will be split in three different categories: prosody speech features hypotheses, voice acoustics features hypotheses and multi-site (UK, Netherlands and Spain) specific hypotheses.

4.2.1. Prosody features

As per observed in the literature, several prosodic features have shown to be associated with depressive symptoms. As a result, and specifically for prosody, the following hypotheses will be investigated:

Hypothesis 1: We have seen in the literature review some conflicting results for the association between longer speech durations with depression severity. Whilst some authors found longer speech durations to be associated with depression (Mundt et al., 2007, 2012), others (Albuquerque et al., 2021) have found duration to be lower. One current hypothesis of this study is that speech duration will be longer for depressed individuals. This is believed to be the case not necessarily because of more vocalization of speech but due to longer pause times, slower speaking rates and lower articulation rates. These additional hypotheses will also be tested in this project (e.g., hypotheses 2, 3 and 4 below).

Hypothesis 2: Huang et al. (2019), Mundt et al. (2007, 2012) and Stasak et al. (2019) highlighted that longer speech durations are believed to be a consequence of longer pause times. The second hypothesis of this study will focus on testing that longer pause times are associated with depression severity.

Hypothesis 3: As per observed in the speech literature, depressed individuals show a slower speaking rate (Cannizzaro et al., 2004; Cummins et al., 2015; Mundt et al., 2007, 2012; Yamamoto et al., 2020). Slower speaking rates are therefore believed to be linked with depression and are believed to be a good predictor for depression severity.

Hypothesis 4: Albuquerque et al. (2021) found slower articulation rate to be a feature that discriminates depressed and non-depressed individuals. It is the project hypothesis that this will also be found in this study.

4.2.2. Voice acoustics features

Hypothesis 5: Mixed results were found in the literature for the role of F0 in discriminating depression. As observed, some authors (Alghowinem et al., 2013; Darby et al., 1984; Low et al., 2020; Mundt et al., 2007) found statistically significant evidence for the role a low F0 plays in discriminating depression, whilst other authors (Mundt et al., 2012; Yang et al., 2013) found no evidence for its role in its association with depression. This project will analyse this further and aims to test the following hypothesis: F0 is a speech feature able to discriminate depressed from non-depressed individuals.

Hypothesis 6: as per the review done by Cummins et al. (2015) loudness has different results in the literature for the association with the severity of depression. In this project the hypothesis is that loudness will be lower for individuals with higher severity of depression.

Hypothesis 7: in the literature review section of this project different results were found for the role speech features derived from read/automated speech and free/spontaneous speech have in discriminating depression (Alghowinem et al., 2013; Mundt et al., 2007). This project sustains this hypothesis and aims to test the hypothesis that voice features derived from free/spontaneous speech and automated/read speech tasks offer different opportunities to discriminate severity of depression.

Hypothesis 8: In a systematic review of using speech to automatically assess several psychiatric disorders**,** Low et al. (2020) reveal more variability on fundamental frequency (F0) is found in depression. As a result, the project hypothesis is that Jitter is higher for depressed individuals and associated with depression**.**

Hypothesis 9: As depressed speech tends to be characterized by having a dull and monotonous tone (Cummins et al., 2015), shimmer (the amplitude of the sound wave) is believed in this project hypothesis to be lower for depressed individuals and associated with depression.

Hypothesis 10: Due to differences in air flow in speech production in depressed individuals (Alghowinem et al., 2013; Teixeira et al., 2013), the current hypothesis is that HNR is lower for depressed individuals and associated with depression.

Hypothesis 11: As per seen in the literature review, speech formants such as F1 and F2 show a difference in depressive speech, potentially due to muscular tension. Although results are not consistent across the literature, it is the aim of this project to test the following hypothesis: Depressive speech is characterised by lower F1 and F2 versus normal speech.

4.2.3. Multi-site nature of study

Hypothesis 12: Because this study also aims to understand if speech voice features are a good basis to predict risk of depression severity, the following will be tested: prosody and voice acoustic features are a good basis to classify depression patients across different geographies.

Hypothesis 13: To the best of the author's knowledge, this is the first longitudinal study of speech features at a large scale and over three different countries. The project aims therefore to test the hypothesis that speech prosody and quality features changes are consistent across cultures in the ability to discriminate changes in depression severity**.**

By exploring these hypotheses this project aims to highlight the value of speech markers as digital phenotypes of depression.

5. Methods

5.1. Overview of Dataset

The dataset used for modelling was from the Remote Assessment of Disease and Relapse in Major Depressive Disorder (Radar-MDD) study, a longitudinal study following 585 patients over 18 months who had a recent history of major depressive disorder (i.e., at least 2 episodes) and with propensity to relapse. These patients have been recruited from clinical cohorts in three countries: UK (King's College London), Spain (Centro de Investigacion Biomedican en Red) and the Netherlands (Netherlands Study of Depression and Anxiety and also patients from Vrije Universiteit Medisch Centrum).

Radar-MDD is part of the Remote Assessment of Disease and Relapse – Central Nervous System (RADAR-CNS) work (Radar-CNS, 2016). Participants in the study wore a wearable fitness tracker (Fitbit charge 2) collecting data on step count, heart rate, sleep patterns, and had additional apps installed for the purpose of the study in their android smartphones. One of the apps was the aRMT (active RMT) collecting data via surveys related with depressive symptoms (via PHQ8).

Alongside this data, every 2 weeks participants completed two speech tasks, where they recorded themselves performing a scripted and an unscripted speech task. The scripted task involved saying aloud in a quiet area an excerpt from 'The North wind and the Sun' (International Phonetic Association, & International Phonetic Association Staff, 1999). Details of the different poem extracts can be found in Appendix I. In the unscripted task participants had to answer the following question: "Can you describe something you are looking forward to in the next week?". The speech data started to be collected in the UK in August 2019 and in Spain and the Netherlands in December 2019.

The present research project focuses on the analysis of these two speech exercises done as part of the Radar-MDD study. The outcome measure to be assessed is depression severity given by PHQ-8 scores (the eight-item Patient Health Questionnaire depression scale) completed at the same time as the speech samples. PHQ8 scores range from 0 to 24. For this project, and whilst building the machine learning classifier, PHQ8 with values of 10 or more were classed as more severe depression, and values of less than 10 as less severe. This cut off value was followed as per guidelines from Kroenke et al. (2009).

Baseline depression and anxiety scores have also been provided by each participant in the three cohorts at the start of the study. Baseline depression score was given by the Inventory of Depressive Symptomatology – Self-Reported questionnaire (Rush et al., 2000). Scores range from 0 to 84 with the following thresholds: 0–13 (no depression symptoms); 14–25 (mild depression symptoms); 26–38 (moderate depression symptoms); 39–48 (severe symptoms); 49–84 (very severe). Baseline anxiety values have been provided through the GAD7 (i.e., 7-item Generalised Anxiety Disorder) questionnaire (Spitzer et al., 2006). Scores range from 0 to 21 with the following thresholds: 0–4 (minimal anxiety), 5–9 (mild), 10–14 (moderate) and

15–21 (severe). For all cohorts the following data was also available: age, gender, height and years of education.

From the total number of participants registered at the beginning of the study, those that provided less than three speech recordings during the longitudinal study have been excluded from the analysis (both for scripted and unscripted speech tasks).

5.1.1. Ethical approval

Ethical approval was obtained from the Camberwell St. Giles Research Ethics Committee (17/LO/1154) in London, from the Fundacio Sant Joan de Deu Clinical Research Ethics Committee (CI: PIC-128-17) in Barcelona, and from the Medische Ethische Toetsingscommissie VUmc (2018.012–NL63557.029.17) in Amsterdam.

5.2. Speech features

A set of 28 speech features were extracted from the scripted and non-scripted speech recordings using parselmouth software. Table 1 shows all the speech features extracted as part of the project. The selected speech features have been based on the literature review and on the hypotheses to be tested.

Table 1

Overview of speech features and expected result (based on hypotheses)

	Speech feature	Description	Expected result
Prosody	Recording Duration	Total duration of recording	Longer for depressed
	Phonation time	Maximum time in which phonation of a vowel is sustained	Lower for depressed
	Phonation ratio	Phonation ratio of sustained vowels	Lower for depressed
	Number of syllables	Total number of syllables	Lower for depressed

	Number of pauses	Total number of pauses	Higher for depressed
	Speaking rate	Number of utterances per second	Lower for depressed
	Articulation rate	Number of words per second	Lower for depressed
	Mean length run	Average number of syllables between pauses	Lower for depressed
	Pause frequency	Number of pauses	Higher for depressed
	Average syllable Duration	Mean duration of syllables	Higher for depressed
	Average pause duration	Mean duration of pauses	Higher for depressed
	Fraction unvoiced frames	Unvoiced frames fraction	Higher for depressed
	Number of voice breaks	Number of voice breaks	Higher for depressed
	Degree of Voice breaks	Degree of Voice breaks	Higher for depressed
Voice quality	Pitch (F0) Mean	Fundamental frequency: lowest frequency of the speech signal (pitch)	Lower for depressed
	Pitch (F0) Std. deviation	Measures dispersion in f0 (variance, standard deviation)	Higher for depressed
	Intensity/Loudness	Mean Loudness	Lower for depressed
	HNR (Mean)	Harmonic to noise ratio (ratio between f0 and noise components)	Lower for depressed
	Jitter (Mean)	Captures irregular closure and asymmetric vocal-fold vibration	Higher for depressed
	Shimmer (Mean)	Voice intensity	Lower for depressed
Articulatory features	F1 frequency (Mean)	First peak in the spectrum (associated with vowels)	Lower for depressed
	F1 frequency (Std. dev)	Measures dispersion in F1 (variance, standard deviation)	Higher for depressed
	F1 bandwidth (Mean)	F1 size of frequency range	Lower for depressed
	F1 bandwidth (Std. dev)	Dispersion in F1 bandwidth (variance, standard deviation)	Higher for depressed
	F2 frequency (Mean)	Second peak in the spectrum (also associated with vowels)	Lower for depressed
	F2 frequency (Std. dev)	Measures dispersion in F2 (variance, standard deviation)	Higher for depressed

F2 bandwidth (Mean)	F2 size of frequency range	Lower for depressed
F2 bandwidth (Std. dev)	Dispersion in F2 bandwidth (variance, standard deviation)	Higher for depressed

5.2.1. Pre-processing of Speech features

The 28 speech features had been previously processed as part of the Radar-MDD (Remote Assessment of Disease and Relapse in Major Depressive Disorder) study. No pre-processing of speech features was done in this project.

5.3. Feature engineering

Feature engineering was done in STATA version 16.1 for Mac. New features were derived and used in statistical modelling and during machine learning classification. The new features created were: depression score (as per best practice values in Kroenke et al., 2009): values of <10 were classed as less severe depression; values =>10 were classed as moderate and more severe depression); age category: 18 to 35 years old; 36 to 60 years and above 60 years old; IDS category (thresholds have been followed as per guidelines by Rush et al., 2000) with values between 0–13:no depression symptoms; 14–25: mild depression symptoms; 26–38: moderate symptoms; 39–48: severe symptoms; 49–84: very severe); GAD7 category with the following thresholds: 0–4 (minimal anxiety), 5–9 (mild), 10–14 (moderate) and 15–21 (severe anxiety); years of education (0-10: up to 10 years of education; 11-20: up to 20 years of education; >=21: up to 30 years of education); baseline PHQ8 (taken as the first value of PHQ8 at first recording) and frequency of speech recordings provided (details of cut-off values detailed in Table 2).

5.4. Statistical analysis

The statistical analysis has been performed in STATA. The project started with descriptive statistics to understand the data (Table 6). Due to the skewness of the data towards females, the median and interquartile ranges (IQR) were reported.

Following from the descriptive statistics analysis, the data was standardised to make it easier to compare scores. All independent variables (except for categorical variables) as well as the dependent variable PHQ8 were standardised to a mean of zero and a standard deviation of one.

5.4.1. Logistic regression model to understand if odds of providing data was associated with depression severity

A logistic regression model per cohort and per speech variant (i.e., scripted and non-scripted) was performed in order to understand if the odds of providing data were associated with severity of depression. The outcome or dependent variable for the logistic regression model was frequency of recordings provided along the 18 months of the study. This dependent variable was binary with a value of 1 for more frequent recordings and 0 for less frequent.

For the UK cohort, scripted recordings frequencies from 1 to 16 were classed as less frequent with the value 0; and frequencies above 17 were classified as more frequent with the value 1. The cut off value of 17 was based on the median response for the UK scripted cohort. For the non-scripted cohort, the cut off value was 15 (median as per Table 6 in the results section). Table 2 reveals all cut off values per country and per recording type for this outcome variable.

Table 2

Frequency of speech recordings dependent variable (for logistic regression) cut *off values*

	Scripted	*Unscripted*
United Kingdom cut off values		
Less frequent: 0	1-16	1-14
More frequent: 1	=>17	=>15
Spain cut off values		
Less frequent: 0	1-10	1-8
More frequent: 1	=>11	=>9
Netherlands cut off values		
Less frequent: 0	1-18	1-15
More frequent: 1	=>19	=>16

As predictors or independent variables, age, gender, and number of education years were added as confounders; baseline GAD7, baseline PHQ8 (i.e., the first reported PHQ8 value at the first recording) and baseline IDS have been added as predictors of interest to understand if baseline depressive symptoms were associated with providing speech data more frequently. Odds-ratio, Wald test chi-square (χ^2)and associated p-levels were reported after fitting the model.

5.4.2. Linear mixed effects models to understand changes in prosody and voice acoustics associated with depression severity

Due to the longitudinal nature of the dataset (i.e., same participants giving speech recordings on multiple occasions, each time point being nested within participants), linear mixed effects models (Stata, 2022). were fitted to the data for each of the three cohorts (UK, Netherlands, and Spain) and according to the speech recording (scripted and unscripted).

For each cohort and speech task (scripted and unscripted), two mixed models were developed with random effects included in the models. The first model only included participant id as a random effect (i.e., random intercept). The second model included both participant id and time as random effects. Model 1 was therefore nested in model 2.

For each model, the outcome variable to measure depressive severity was PHQ8 (obtained straight after the recording). Confounders included in the model have been age, height, gender, education years, baseline IDS, baseline GAD7. The 28 speech features of Table 1 were fitted as independent variables.

When fitting the linear mixed effects models, the maximum likelihood (MLE) fitting routing was adopted. MLE has proved to be effective in the presence of missing data when data is assumed to be missing at random. For each model, z-statistics per speech feature were reported alongside the respective p-value. Likelihood ratio tests were performed to compare models with participant id and time as random effects, versus the nested model with participant id only.

5.5. Machine learning classifiers to predict severity of depression based on speech features

The output from the STATA data cleaning and feature engineering stages for all cohorts and speech tasks was then used as the dataset for the machine learning classifiers. The focus of the machine learning classifiers was on the UK and Dutch cohorts. By running the STATA do files provided with this research project for the UK and Dutch cohorts, it is possible to extract each of the datasets used for the classifiers per country and speech task to replicate the results. Instructions are provided in the do files comments.

To answer the second reseach question (i.e., can speech features be used to classify depression severity) classification was done by fitting several classifiers to the data. Three modelling techniques widely used in the literature for speech analysis (Cummins et al., 2015) have been Support Vector machines (SVM), Gaussian Mixture models (GSM) and Random Forest. In this project the following machine learning techniques will be explored: SVMs, Random Forest, XGBoost, Feedforward Neural Networks and Convolutional Neural Networks.

To develop the classifiers, the anaconda environment was used with python version 3.7. The data was split into a training dataset and a testing dataset. The split, stratified by participant, was 70/30 i.e., 70% of the data was used for training and 30% to test the classifiers.

The classifiers were built per cohort and speech task. In the first set of experiments the classifiers for all the cohorts and speech tasks, had as independent variables or predictor variables only the 28 speech features. As outcome or dependent variable of interest the classifiers had to predict the binary depression score measuring severity of depression (PHQ8 scores <10 were classed as less severe with a 0; PHQ8 scores =>10 as more severe and classed as 1). The objective was to develop a classifier that performed well by predicting the outcome variable (y) in unseen data.

In the second set of experiments, the classifiers for all the cohorts and speech tasks, included the 28 speech features as independent variables and in addition also included demographics (age, gender, height and education years), baseline depression (baseline PHQ8, baseline IDS) and anxiety scores (baseline GAD7).

Training data was used for the classifiers to learn via 5-fold GroupTimeSeries split cross validation (Figure 1) to find the optimal parameters to predict severity of depression (i.e., binary classification: PHQ8 <10 as 0; 1 meaning more severe for PHQ8 values of =>10). This cross validation technique used in this project was based on the code for a feature request waiting to be released to scikit-learn (Chawla, 2019 & Kaggle, 2020). This ensured that the cross validation kept both the participants and time series in the correct order during cross validation with no data leakage.

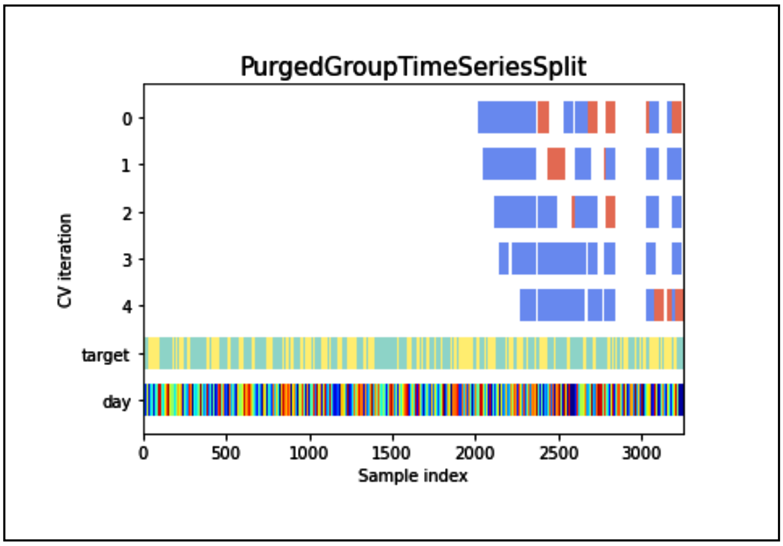

Figure 1. *Group Time series split for cross validation.*

The final testing phase allowed to test the classifiers performance on unseen data to ensure external validity.

The y labels (i.e., depression score variable) were checked for data imbalance. Because the patients recruited for this study have suffered from MDD in the previous 2 years, the classes were not very imbalanced (appendix III).

The independent variables were also normalised using min-max normalisation via the MinMaxScaler pre-processing scikit-learn tool. This normalisation performs a linear transformation of the data into a range of (0,1).

When building the classifiers, this project started by developing SVM classifiers using the scikit-learn library. A total of 5 baseline SVM classifiers were developed (Table 3). As mentioned, Group Time series split cross validation and grid search were used to find the optimal parameters for complexity (C), gamma and kernel. The chosen metric to assess performance was ROC AUC.

Table 3

SVM Grid search values to find optimal hyperparameters

	Parameters		
	C	*gamma*	*kernel*
SVM 1	$[1^{-5}, 1^{-4}, 1^{-3}, 1^{-2}, 1^{-1}, 1, 10, 100]$	$[1^{-5}, 1^{-4}, 1^{-3}, 1^{-2}, 1^{-1}, 1, 10, 100]$	[RBF, sigmoid, linear, poly]
SVM 2	[50, 10, 1, 0.1, 0.01]	[50, 10, 1, 0.1, 0.01]	[RBF, sigmoid, linear, poly]
SVM 3	[50, 10, 1, 0.1, 0.01]	scale	[RBF, sigmoid, linear, poly]
SVM 4	5	scale	RBF
SVM 5	1	1	RBF

A Random Forest classifier was developed following hyperparameters detailed in Table 4. Like the SVM classifiers, Group Time series split cross validation and grid search were used to find the optimal parameters. The chosen metric to evaluate the performance of this classifier and to allow comparison between all the models in this project was also ROC AUC.

To build on the classifiers developed during the experimentation stage, an XGBoost was also developed. The hyperparameters considered when developing the XGBoost model can be found in Table 4. These hyperparameters have been chosen after reading several sources (XGBoost developers, 2021; Zhang, 2015). A similar approach to the previous models has been followed for cross validation and performance measure chosen to assess the classifier.

Table 4

Random Forest and XGBoost grid search values to find optimal hyperparameters

	Random Forest	*XGBoost*
Number of trees	[100, 200, 300, 1000]	[100, 200, 300, 500, 1000]
Maximum depth of the trees	[80, 90, 100, 110]	[4, 6, 8, 10]
Learning rate		[0.0001, 0.001, 0.01, 0.1]
Features to consider when looking for the best split	[2, 3]	
Samples required to be at a leaf node	[3, 4, 5],	
Minimum samples required to split the node	[8, 10, 12]	

The last part of the experiments focused on testing deep learning models. A total of five neural networks were built (Table 5). One feedforward dense neural network and four convolutional neural networks.

The feedforward dense neural network had two hidden layers. The hidden layers had a relu activation function and the final layer had a softmax function to perform the final classification. The Adam optimiser was used to compile the model, with an initial learning rate of 1^{-4} and a binary cross entropy loss function. The layer's bias and kernel used ridge regularisation with a value of 0.01. This classifier was trained with 50 epochs and a batch size of 10.

The four convolutional neural networks had similar settings with the exception for the number of filters used in the hidden layers and the optimiser used. The first CNN model, model 1, had two hidden layers and batch normalisation. The activation function in the hidden layers was a relu (similar setting to the feed forward dense neural network) with a final softmax layer to perform the final activation. The SGD optimiser was used to compile the model with an initial learning rate of 1^{-4} and a binary cross entropy loss function. This classifier was trained with 64 epochs and a batch size of 32. CNN Model 2 had a similar setting to CNN Model 1 with the only difference being the optimiser, which was changed from SGD to the Adam optimiser, all other setting remained the same.

Model 3 CNN kept the same set up for the number of hidden layers and batch normalisation but had an increase in terms of the number of filters on the hidden layers and the optimiser was changed back to the SGD one. All the other settings remained the same. The last CNN model, Model 4 CNN, had only one hidden layer with an increase in the number of filters at this hidden layer. All the other settings, i.e., batch normalisation, activation functions, optimiser, loss function were kept the same as model CNN 3.

Table 5

Deep Neural Networks settings

	Hidden layers	*Activation hidden /final*	*Optimiser*	*Loss function*
FF Neural Network	2	Relu/softmax	Adam	Cross entropy
CNN 1	2	Relu/softmax	SGD	Cross entropy
CNN 2	2	Relu/softmax	Adam	Cross entropy
CNN 3	2	Relu/softmax	SGD	Cross entropy
CNN 4	1	Relu/softmax	SGD	Cross entropy

In the results section the performance will be reported for of all these 12 classifiers. Results will be provided for the UK and Netherlands cohorts for the scripted and free speech tasks. Therefore, a total of 24 classifiers (12 classifiers with only the speech features as predictors + 12 classifiers with the speech features plus demographics and baseline depression and anxiety predictors) per task (scripted and free speech) and per cohort will be reported. The reported results will therefore be for a total of 96 classifiers.

In addition, LIME (Local Interpretable Model-agnostic Explanations) will be used to derive explanations of the features that are contributing for the classification. This is an off the shelf algorithm to explain the reasons behind a specific prediction. The advantage of using this tool is that it can be applied to any classifier (Ribeiro et al., 2016). LIME uses a local linear approximation of a specific instance via applying perturbation around it which serves as an explanation for the classification.

In the results section, the best performing model for each cohort and speech task will be analysed in detail for the interpretability behind the classification.

5.6. Missing data

During statistical analysis, and due to the results of the logistic regression model on the odds of providing recordings to be associated with depression severity being minimal, missing data was treated as missing at random and included in the linear mixed effects modelling. During machine learning classifier development, missing data was imputed via multiple imputation using the mean of the respective variable (via the Scikit-learn library SimpleImputer function).

6. Results

6.1. Descriptive statistics

As per Table 6, there is a difference in the median of speech recordings per site, with participants in the UK (median: 17 for scripted; 15 for unscripted) and Netherlands (median: 19 for scripted; 16 for unscripted) providing more speech recordings than the Spanish cohort (median: 11 and 9 respectively). The cohort with the highest number of participants was the UK one with a total of 248 patients providing at least 3 samples of scripted speech and 240 patients providing at least 3 samples of unscripted speech recordings (Table 6). In terms of gender, across the three cohorts there was a skew towards females (Table 6). Across scripted and non-scripted speech recordings, 77% participants were female in the UK, a total of 67% participants in Spain were female and a total of 79% participants in the Netherlands were female.

As per Table 6, the UK and Netherlands cohorts had a similar age profile and were slightly younger than the Spanish cohort. As for years of education, patients in the UK and Netherlands cohorts had similar number of years of education, with a median of 17 years, a higher number versus the Spanish cohort.

Both baseline IDS and baseline GAD7 were higher for the Spanish cohort, with a median of 38 and 12 respectively. This suggest that the Spanish cohort had a higher severity of baseline depressive and anxiety symptoms versus the British and the Dutch cohorts.

When looking at the outcome variable of interest, PHQ8, it can be observed that the Spanish cohort had higher levels of depression versus the other two (Figure 2).

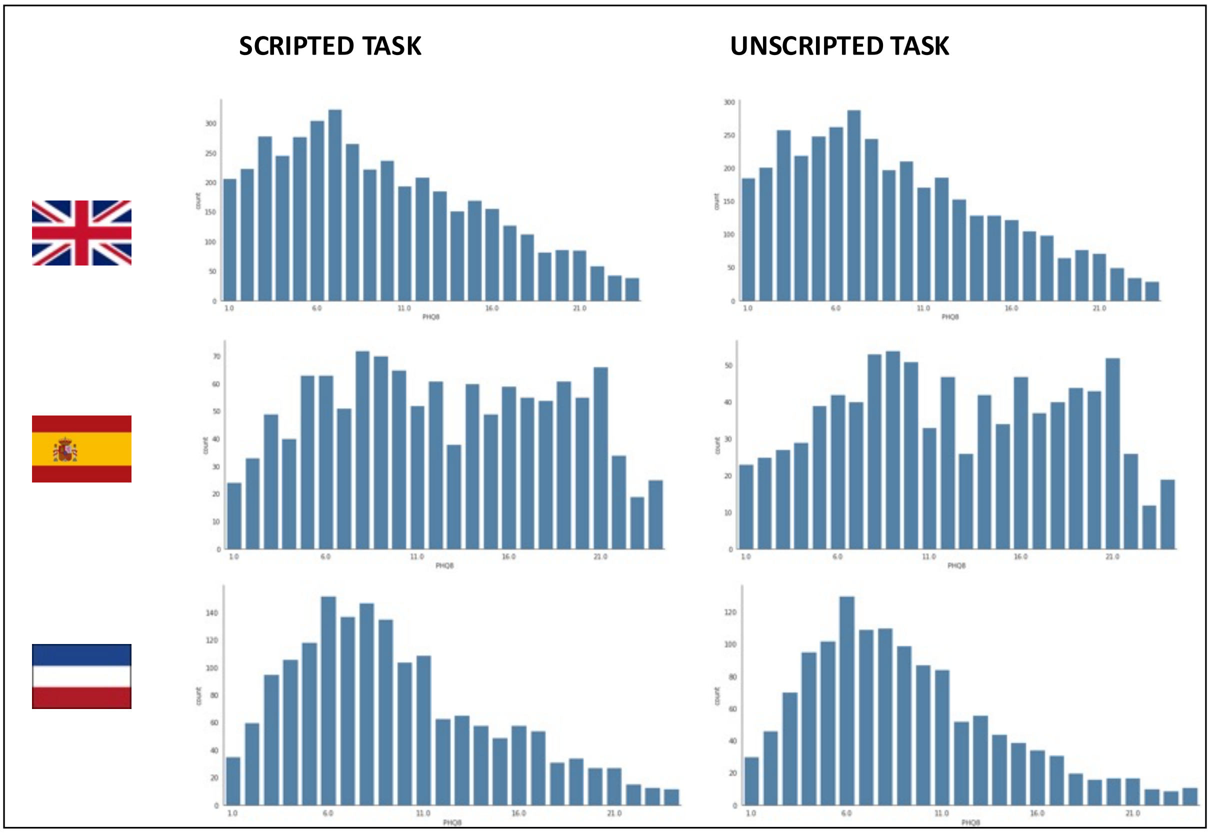

Figure 2. *Distribution of PHQ8 per cohort and speech task*

Table 6. *Overview of descriptive statistics for the three cohorts*

		United Kingdom		Spain		Netherlands	
		Scripted	Unscripted	Scripted	Unscripted	Scripted	Unscripted
	n (=>3 speech samples)	**248 (273 total)**	**240 (267 total)**	**91 (115 total)**	**77 (110 total)**	**98 (107 total)**	**92 (105 total)**
Gender	Female	192	185	62	51	78	73
	Male	56	55	29	26	20	19
Age at start of study	Median	46	47	53	55	43.5	47.5
	IQR	[31-59]	[31-60.5]	[46-60]	[48-60]	[26-58]	[26-58.5]
Education years	Median	17	17	13	12	17	17
	IQR	[14-19]	[14-19]	[11-17]	[10-16]	[14-20]	[14-21]
Baseline IDS	Median	27	27	38	38	29.5	29
	IQR	[20-37]	[19-36.5]	[28-52]	[28-52]	[20-38]	[20-37]
Baseline GAD7	Median	7	7	12	12	7	7
	IQR	[4-12]	[4-12]	[7-16]	[7-16]	[4-10]	[4-10]
# of recordings	Median	17	15	11	9	19	16
	IQR	[9-26]	[8-22]	[5-21]	[5-17]	[11-24]	[7-22]

6.2. Odds of providing more frequent speech recordings per depression severity and demographics

As per Table 7, after fitting the logistic regression model, no significance was found amongst the baseline depression (baseline IDS and baseline PHQ8) and anxiety (GAD7) predictors for the UK and the Spanish cohorts as for the odds of providing more frequent speech recordings. Looking at the other confounding variables fitted for these two cohorts, and in the specific case of the UK cohort, age (scripted: z=3.72, p<.001, odds ratio = 1.76; unscripted: z=2.37, p<.05, odds ratio = 1.42) and gender (scripted: z=-1.99, p<.05, odds ratio = 0.50, unscripted: z=-2.45, p<.05, odds ratio = 0.43) were the only variables significantly associated with the number of scripted and spontaneous speech recording samples provided. These results suggest that the odds ratio of providing more recordings increases per each one-year increase in age in this cohort. It was also observed for the UK cohort that gender plays a role. Being a male sees a decrease by 50% (for scripted speech) and by 57% (unscripted) on the odds of providing more frequent speech recordings versus females.

Baseline PHQ8 in the Dutch cohort for spontaneous speech is the only significant predictor variable significantly associated with the odds of providing speech recordings more frequently. For each unit increase of baseline PHQ8 a 54% (1-0.46 =0.54) decrease in the odds of providing speech recordings is observed for unscripted recordings. No similar finding was observed for the scripted tasks, nor in the other two cohorts for any of the recording tasks.

Table 7. *Logistic regression results assessing if number of speech recordings was affected by baseline depressive and anxiety symptoms*

		UK		**Spain**		**Netherlands**	
		Odds ratio	Z statistic	Odds ratio	Z statistic	Odds ratio	Z statistic
Age	Scripted	**1.76**	**3.72*****	0.74	-1.09	1.29	1.02
	Unscripted	**1.42**	**2.37***	0.52	-2	1.39	1.25
Male	Scripted	**0.50**	**-1.99 ***	2.03	1.28	1.08	0.13
	Unscripted	**0.43**	**-2.45***	1.50	0.64	2.11	1.19
Education years	Scripted	0.85	-1.08	1.13	0.43	0.95	-0.21
	Unscripted	1.08	0.46	0.86	-0.53	0.96	-0.17
Baseline GAD7	Scripted	0.90	-0.53	1.48	0.87	0.78	-0.9
	Unscripted	0.87	-0.69	1.12	0.26	0.62	-1.67
Baseline PHQ8	Scripted	1.10	0.4	0.51	-1.73	0.67	-1.4
	Unscripted	0.91	-0.4	1.26	0.46	**0.46**	**-2.37***
Baseline IDS	Scripted	0.74	-1.19	0.89	-0.28	1.02	0.56
	Unscripted	1	0.03	0.49	-1.16	1.61	1.34

Note. * indicates p < .05, ** indicates p < .01, *** indicates p < . 001

6.3. Prosody and voice quality features associated with depression severity

For each of the cohorts, two mixed effects linear models were fitted to the data. Except for the Spanish cohort, the models including participant id and time as random effects for the UK and Dutch cohorts fitted the data better for both speech tasks. In the case of the Spanish cohort, linear mixed effects models with just participant id as random effect fitted the data better than including participant id and time as random effects.

6.3.1. UK cohort results: scripted task

Likelihood ratio test revealed that including both a random intercept and slope fitted the data better: χ^2 (1) =333.97, *p*=0.000. Therefore, the reported results in this section are for this model. The results (Table 8) for the UK cohort reveal 5 out of the 28 speech features to be associated with depression severity. Out of the prosody speech features, recording duration, number of pauses, average number of syllables between pauses and loudness are significantly associated with depression. Starting with recording duration, the relationship with severity of depression is negative, in the sense that for a unit increase of recording time, there is a decrease of 1.20 (*p*=0.027) in PHQ8. Although significance was only seen for the UK cohort in the scripted task, the relationship had the same trend for the two other cohorts and across both speech tasks albeit no significant. This suggests that longer recordings are not associated with depression severity.

Another prosody feature significantly associated with depression severity is the number of pauses during speech, with an increase of 0.67 in PHQ8 (*p*=0.03) for every extra unit of speech pause. The results for mean length run (average number of syllables between pauses) indicate that for each additional syllable between pauses there is an increase of 0.33 (p=0.003) in depression severity; and for intensity mean (loudness) an inverse relationship is observed where for every unit increase of loudness there is a 0.29 (p=0.000) decrease in depression severity. The voice quality feature significantly associated with severity of depression was HNR. Results show for every one unit increase of HNR (*p*=0.013) there is a 0.32 increase in PHQ8.

6.3.2. UK cohort results: unscripted task

Like the scripted task results, likelihood ratio test revealed that including both a random intercept and slope fitted the data better also for the unscripted task: χ^2 (1) =303.48, *p*=0.000.

Only one speech feature on the unscripted task is significantly associated with depression severity. This is intensity mean (loudness), revealing the same inverse trend seen in the scripted task. For every unit increase of loudness there is a 0.29 (p=0.004) decrease in depression severity. No other unscripted speech features were significantly associated with depression severity.

6.3.3. Spanish cohort results: scripted task

Contrary to the results of the UK cohort, in the case of the Spanish cohort, likelihood ratio test revealed that including just the random effect for participant id fitted the data better, as opposed to including both participant id and time: χ^2 (1) =0.62, *p*=0.43. The reported results in this section are for this model, i.e., including just participant id as random effect.

Four speech features during the scripted task were significantly associated with depression severity amongst the Spanish cohort. Three are prosody features and one is articulatory feature. However, none of the features are the same as the observed for the UK cohort. Specifically on the prosody features, phonation ratio (-0.30, p=0.025) is associated with depression severity. The relationship is negative, where for each unit increase of this ratio, there is a 0.30 decrease in severity of depression. Speaking rate (0.33, p=0.049) and articulatory rate (-0.33, p=0.009) are also associated with severity of depression severity. Revealing that an increase in speaking rate is associated with a more severe depression, and an increase in articulatory rate is associated with less severe or no depression.

F1 frequency mean (-0.09, p=0.009) was the formant feature associated with depression severity. An increase in F1 frequency mean reveals a decrease in PHQ8 score and therefore less severity of depression.

6.3.4. Spanish cohort results: unscripted task

In line with the results for the scripted task, likelihood ratio test for the unscripted task of the Spanish cohort, revealed that including just the random effect for participant id fitted the data better, as opposed to include both participant id and time: χ^2 (1) =3.24, *p*=0.07. Fitting this model, two speech articulatory features revealed to be significantly associated with depression. F1 bandwidth (mean) revealed that for a one unit increase of F1 bandwidth, there is a 0.07 (*p*=0.034) increase in PHQ8; F1 frequency mean (-0.09, p=0.005) was also associated with severity of depression.

6.3.5. Dutch cohort results: scripted task

Like the UK results, likelihood ratio test revealed that including both a random intercept and slope fitted the data better also for the Dutch cohort: χ^2 (1) =97.18, *p*=0.000. Out of the three cohorts, the Dutch cohort saw the highest number of speech features being associated with changes in depression severity. A total of six speech features during the scripted task were found to be significant. Number of pauses (0.18, p=0.031) and pause frequency (-0.21, p=0.005) have been the prosody speech features associated with depression. On measures of voice quality, pitch (F0) mean (-0.17, p=0.002), HNR (0.13, p=0.004) and Jitter (0.056, *p*=0.05) were associated with depression. Both HNR and Jitter results revealed that for an increase in HNR and Jitter, PHQ8 scores would also increase and as a consequence depression severity. HNR results are in line with the results found in the UK cohort for the scripted task, i.e., higher levels corresponding to more severe depression.

F1 frequency mean (0.07, *p*=0.008) was the only articulatory speech feature associated with depression for the scripted task in the Dutch cohort. Unlike the UK cohort, the relationship with PHQ8 was positive. This means that for the Dutch cohort and for scripted tasks, increases in F1 frequency (mean) are associated with increases of depression scores.

6.3.6. Dutch cohort results: unscripted task

Likelihood ratio test revealed that including both a random intercept and slope fitted the data better for the unscripted task: χ^2 (1) =50.34, *p*=0.000. The reported results in this section are for this model. Pause frequency (-0.08, *p*=0.042), pitch mean (-0.16, *p*=0.001) and intensity mean (-0.10, *p*=0.028) have been the speech prosody features associated with depression for the unscripted task in the Dutch cohort.

HNR (0.09, *p*=0.017) was the only measure of voice quality for the unscripted task associated with severity of depression with significance. Like the scripted task, higher levels of HNR correspond to higher levels of depression severity.

On measures of articulatory speech, F1 bandwidth standard deviation (-0.06, *p*=0.037) was found to be significant for the association with depression severity. The relationship is negative meaning that higher values of F1 bandwidth variation correspond to less severe depression.

Table 8. *Linear Mixed effects models results for the association between speech features and PHQ8 (severity of depression)*

		UK		Spain		Netherlands	
		Coefficient	Z statistic	Coefficient	Z statistic	Coefficient	Z statistic
Prosodic features							
Recording Duration	Scripted	**-1.20**	**-2.21***	-0.06	-0.8	-0.10	-0.73
	Non scripted	-0.15	-1.44	-0.08	-0.65	-0.20	-1.33
Phonation time	Scripted	0.52	1.08	0.10	0.68	0.14	0.66
	Non scripted	-0.04	-0.4	-0.008	-0.04	0.11	0.48
Phonation ratio	Scripted	-0.52	-1.12	**-0.30**	**-2.24***	-0.07	-0.41
	Non scripted	-0.01	-0.24	-0.16	-1.73	-0.13	-1.11
Number of syllables	Scripted	0.51	1.42	-0.03	-0.31	-0.04	-0.25
	Non scripted	0.14	1.48	0.07	0.38	0.20	1.01
Number of pauses	Scripted	**0.67**	**2.17***	-0.06	-0.77	**0.18**	**2.15***
	Non scripted	0.06	1.25	0.02	0.28	0.07	0.89
Speaking rate	Scripted	-0.17	-0.28	**0.33**	**1.97***	-0.096	-0.45
	Non scripted	-0.07	-1.21	0.07	0.54	-0.01	-0.09

Articulation rate	Scripted	-0.24	-0.6	**-0.32**	**-2.6****	0.01	0.07
	Non scripted	0.01	0.32	0.008	0.12	-0.00	-0.01
Mean length run	Scripted	**0.33**	**2.95****	-0.03	-1.11	0.02	0.62
	Non scripted	0.006	0.43	0.025	0.84	0.01	0.38
Pause frequency	Scripted	-0.31	-1.27	-0.01	-0.23	**-0.21**	**-2.8****
	Non scripted	-0.02	-0.94	0.01	0.28	**-0.09**	**-2.04***
Average syllable duration	Scripted	0.01	0.22	-0.05	-1.6	0.004	0.06
	Non scripted	0.02	1.3	0.07	1.54	0.03	0.83
Average pause duration	Scripted	0.09	1.08	-0.002	-0.06	-0.03	-1.26
	Non scripted	-0.001	-0.11	-0.01	-0.69	0.01	0.52
Unvoiced framed fraction	Scripted	-0.21	-1.67	0.001	0.03	-0.04	-0.8
	Non scripted	-0.004	-0.19	-0.07	-1.19	-0.03	-0.58
Number voice breaks	Scripted	-0.13	-1.13	0.04	1.27	-0.02	-0.57
	Non scripted	-0.02	-0.28	-0.03	-0.47	-0.15	-1.49
Degree voice breaks	Scripted	0.21	1.89	0.01	0.32	-0.03	-0.72
	Non scripted	0.029	1.64	0.04	1.23	0.06	1.58

Voice quality features							
Pitch (F0) mean	Scripted	-0.29	-1.9	0.03	0.86	**-0.17**	**-3.5****
	Non scripted	-0.02	-0.81	-0.02	-0.59	**-0.16**	**-3.6****
Pitch (F0) Std Deviation	Scripted	0.04	0.52	0.008	0.35	0.01	0.53
	Non scripted	0.001	0.09	0.009	0.36	-0.02	-0.84
Intensity (loudness) mean	Scripted	**-0.41**	**-3.6*****	-0.01	-0.44	-0.03	-0.7
	Non scripted	**-0.06**	**-2.91****	0.03	0.78	**-0.10**	**-2.19***
HNR	Scripted	**0.32**	**2.47***	0.02	0.41	**0.13**	**2.89****
	Non scripted	0.002	0.11	0.006	0.14	**0.09**	**2.39****
Jitter	Scripted	0.03	0.37	0.025	0.83	**0.056**	**1.96***
	Non scripted	0.03	1.77	0.04	1.36	0.04	1.21
Shimmer	Scripted	-0.02	-0.23	0.009	0.28	-0.04	-1.28
	Non scripted	0.01	0.72	0.02	0.53	-0.03	-0.73
Articulatory features							
F1 frequency (mean)	Scripted	0.002	0.03	**-0.06**	**-2.09***	**0.075**	**2.67****
	Non scripted	-0.008	-0.62	**-0.09**	**-2.81****	0.06	1.79

F1 frequency (Std. Dev)	Scripted	-0.01	-0.15	-0.03	-1.12	-0.03	-1.14
	Non scripted	0.01	0.97	-0.02	-0.68	-0.02	-0.65
F1 bandwidth (mean)	Scripted	0.11	0.87	0.07	1.67	-0.02	-0.61
	Non scripted	-0.01	-0.53	**0.07**	**2.12***	0.06	1.54
F1 bandwidth (Std Dev)	Scripted	-0.07	-0.73	0.008	0.26	0.02	0.89
	Non scripted	0.003	0.21	-0.01	-0.37	**-0.06**	**-2.09***
F2 frequency (mean)	Scripted	-0.05	-0.61	0.02	0.99	0.006	0.24
	Non scripted	0.004	0.36	-0.01	-0.46	0.024	0.91
F2 frequency (Std. Dev)	Scripted	-0.05	-0.74	0.04	1.81	-0.003	-0.15
	Non scripted	-0.002	-0.2	0.04	1.48	0.03	1.33
F2 bandwidth (mean)	Scripted	0.03	0.24	-0.004	-0.12	-0.03	-0.8
	Non scripted	0.0006	0.04	-0.01	-0.3	-0.01	-0.38
F2 bandwidth (Std Dev)	Scripted	-0.06	-0.66	0.002	0.06	-0.04	-1.26
	Non scripted	0.006	0.44	0.001	0.04	0.01	0.44

Note. * indicates p < .05, ** indicates p < .01, *** indicates p < .001

6.4. Predicting severity of depression via machine learning classifiers

The reported results in this section are for the UK and Dutch cohorts. The rational to focus on these cohorts only for the machine learning classifiers was a) due to the number of participants being higher in the UK cohort (this cohort was the highest in terms of number of participants amongst the three cohorts); b) in the case of the Dutch cohort it was the one with highest number of speech features being significantly associated with depression severity when fitting the linear mixed effects models. The metric chosen to evaluate the machine learning classifiers for the two cohorts was ROC-AUC.

6.4.1. UK cohort results: scripted task

A total of 24 classifiers were built for the scripted speech task. As per Figure 3, it can be observed across all the classifiers (with the exception for the SVM baseline classifier), that including demographics, baseline anxiety and depression values with the 28 speech features, derives better AUC performance than speech features only.

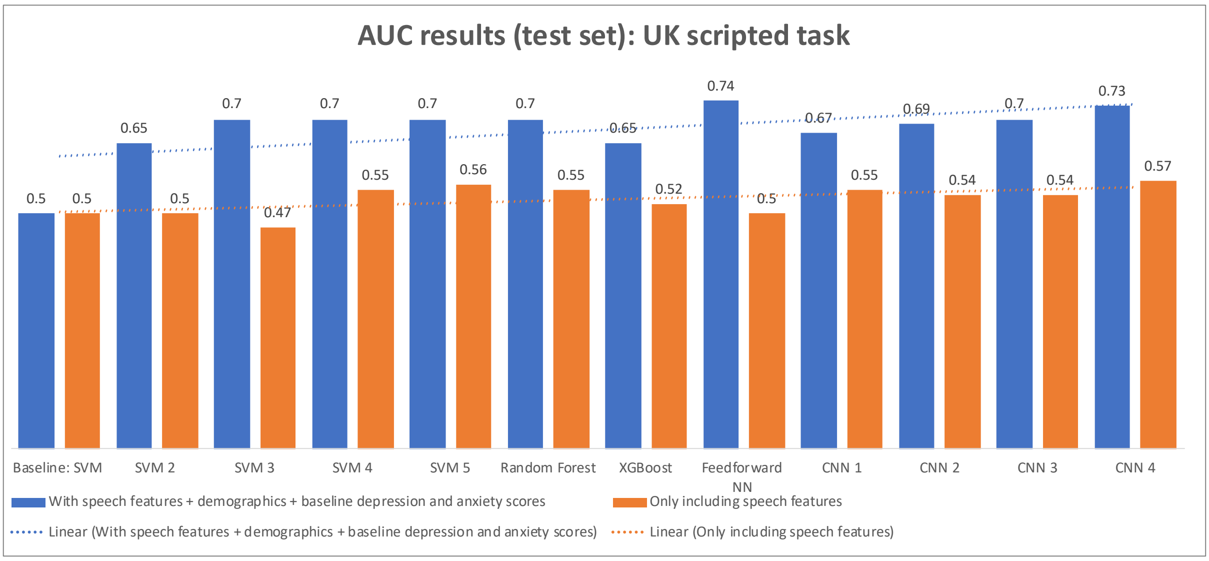

Figure 3. *UK AUC results on test dataset for scripted task*

Out of these classifiers, the machine learning model that perform the best in predicting severity of depression for the UK scripted task, was the feedforward dense neural network (model 8 for the scripted task). As per Table 5, this network had a relu activation function at the hidden layers and a final layer with a softmax function to perform the final classification. To compile the model, the Adam optimiser was used with an initial learning rate of 1^{-4} and a binary cross entropy loss function. The layer's bias and kernel used ridge regularisation with a value of 0.01.

With a very similar performance in terms of AUC was the fourth convolutional neural network including just one hidden layer with an increase in the number of filters at this hidden layer. Both these models had an AUC above 70 which proves the classifier's ability to discriminate well between negative and positive cases.

Looking at the results of the 12 classifiers (Figure 3) that only included the 28 speech features as independent variables, their performance was not good. The AUC was always below 0.60 which reveals the models only with speech features as predictors performed at chance level.

6.4.2. UK cohort results: unscripted task

The results for the classifiers when predicting severity of depression from free speech tasks, reveal a similar pattern. The 12 classifiers that include the 28 speech features together with demographics, baseline depression and anxiety variables perform better in classifying depression severity versus those that only include the speech features (Figure 4).

Out of these 12 classifiers, and like observed for the scripted task, the classifier that performed the best was the feedforward dense neural network. This network had the same set up as the feedforward neural network developed for the scripted task activity and observed in Figure 3.

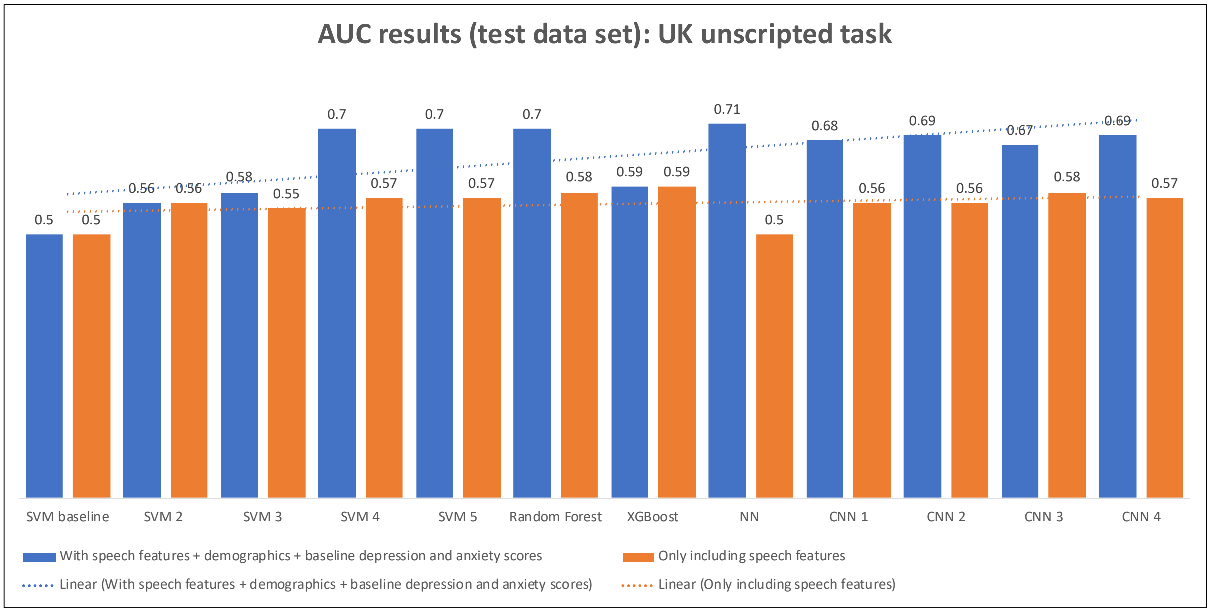

Figure 4. *UK AUC results on test dataset for unscripted task*

Of importance to note also is the result of the SVM 4, SVM 5 and the Random Forest model which all have an AUC of 0.70, also revealing robust results in discriminating negative and positive cases.

Once again, the results of the 12 classifiers only using speech features as independent variables, did not perform robustly in predicting severity of depression. All these classifiers have an AUC ranging between 0.5 and 0.59, which does not discriminate well less severe more severe cases.

6.4.3. Dutch cohort results: scripted task

The same trend was observed for the need to include demographics, baseline anxiety and depression variables on the classifiers alongside the 28 speech features, as these are the ones performing the best also in the Dutch cohort in classifying severity of depression. However, when comparing the results between the Dutch and the UK cohorts, the classifiers in the Dutch cohort achieve a lower AUC performance (Figure 5).

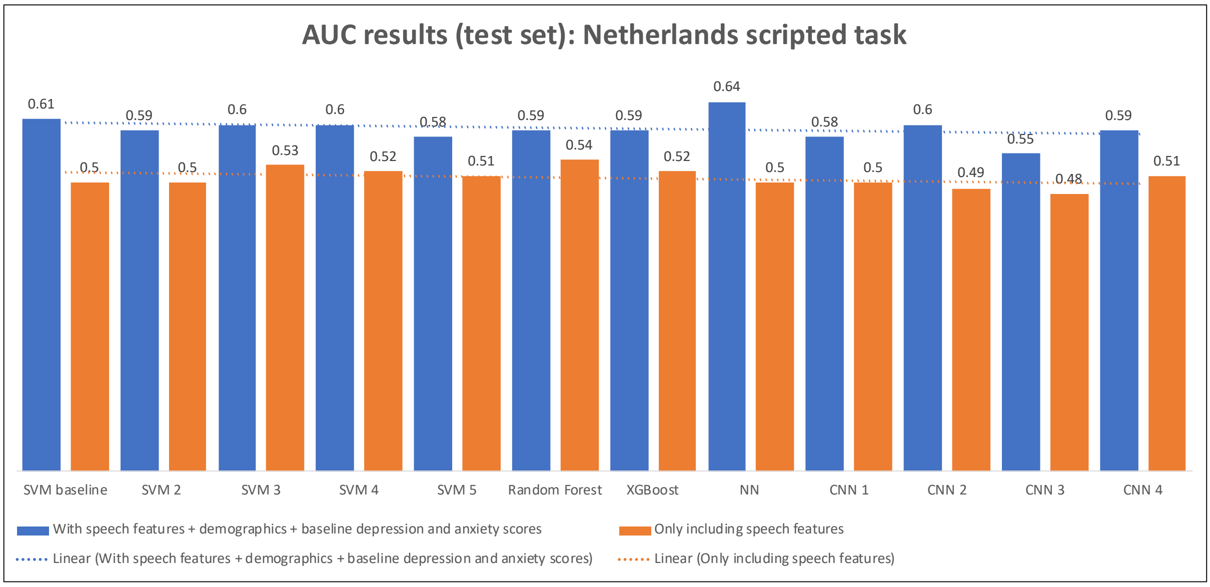

Figure 5. *Dutch AUC results on test dataset for scripted task*

When predicting severity of depression during a scripted task, it can be observed the best classifier is still the feed forward dense neural network, as per also observed in the UK scripted and spontaneous task, albeit with a lower AUC performance (0.64 in the scripted Dutch cohort versus 0.74 in the UK scripted cohort). The classifiers built to predict depression severity from speech only variables during the scripted task did not offer robust results. Again, AUC for these classifiers is low, with values ranging between 0.5 and 0.54.

6.4.4. Dutch cohort results: unscripted task

Same trends for best results when including speech features, demographics and baseline anxiety and depression as independent variables to predict severity of depression. It can be seen the classifier that perform the best for this cohort during unscripted tasks, was the Random Forest and not the Feedforward neural network (Figure 6). As per before, none of the classifiers including only speech features as independent variables had an acceptable AUC performance.

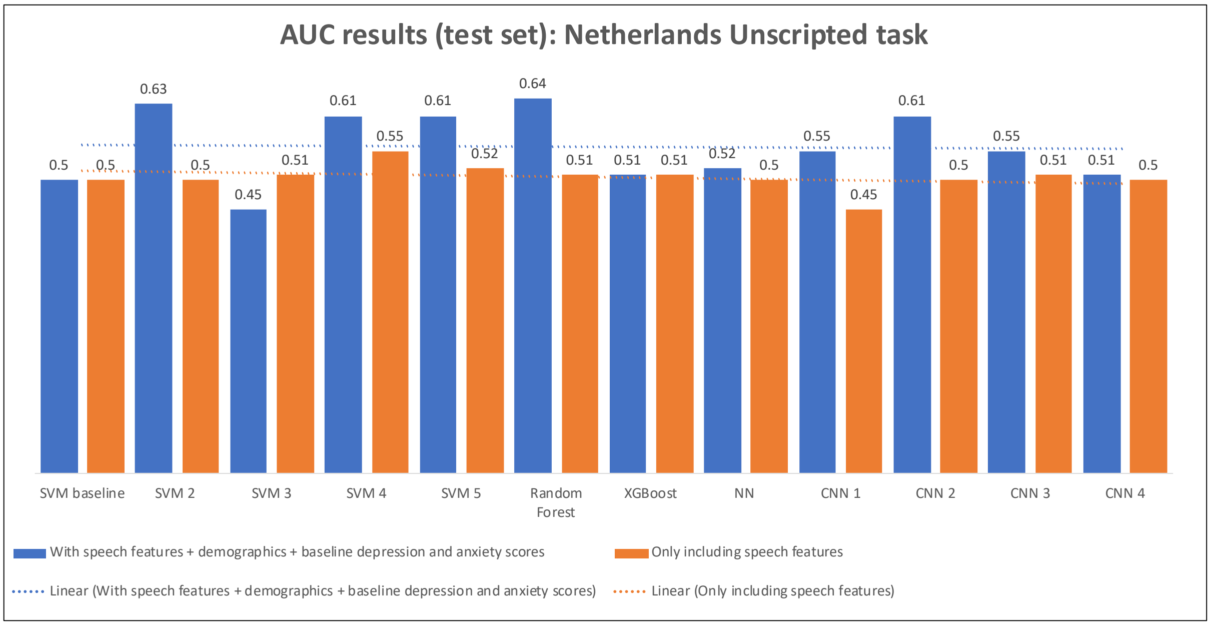

Figure 6. *Dutch AUC results on test dataset for unscripted task*

6.4.5. Interpretability behind the best AUC classifiers

Using LIME (Local Interpretable Model-agnostic Explanations) authored by Ribeiro et al. (2016) it is possible to understand the independent variables that are contributing for True Positives, False Positives, True Negatives and False Negatives. In this section the focus will be in understanding the variables behind the binary classification, i.e., a less severe or a more severe depression case, from the top classifiers per cohort and speech task.

More precisely, the best performer classifiers for the UK and Dutch cohorts per speech task will be investigated for the features that are contributing to a specific outcome. For the UK cohort, these are the Feedforward Neural Networks for scripted and unscripted speech tasks. For the Dutch cohort, the Feedforward Neural Network in the case of the scripted task and the Random Forest classifier in the case of the unscripted task. Table 9 provides an overview of the classifiers.

Table 9

Top Performer Classifiers being interpreted with LIME

	UK cohort	Netherlands cohort
Scripted task	Feedforward Neural Network	Feedforward Neural Network
AUC	0.74	0.64
Unscripted task	Feedforward Neural Network	Random Forest
AUC	0.71	0.64

6.4.5.1. UK cohort

Looking at Figure 7 it can be observed that certain features are contributing for the prediction of either a less severe or more severe depression case when classifying depression severity from scripted speech tasks. The classifier is 64% confident that this is a severe depression case. A true positive case, or more severe case, is being derived from high baseline IDS, higher F2 bandwidth variability, lower mean pause duration and to a lesser extent by higher values of fraction of unvoiced frames, number of pauses and F2 frequency variability.

Of importance to highlight is the role F2 bandwidth variability (i.e., F2 bandwidth standard deviation) and baseline IDS play in classifying depression severity during scripted tasks. High values of baseline IDS and variability in F2 bandwidth are key variables in contributing for the chance of an instance to be classified as severe. Of interest to point out is the values for average pause duration, where lower values for this variable increase the chance of a case to be classed as severe (a lower number increases the chance of being classified as a severe case).

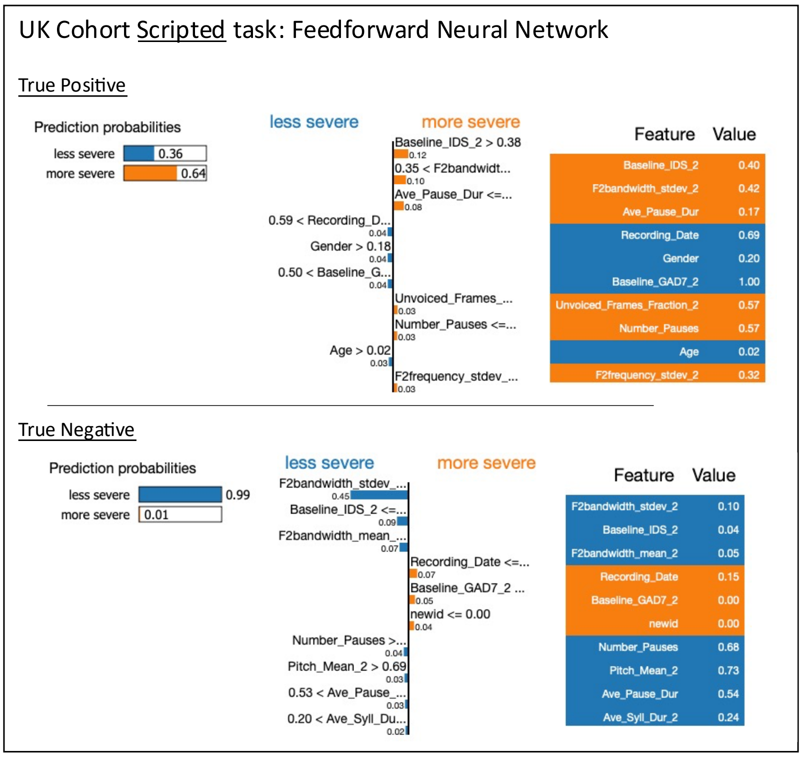

Figure 7. *UK cohort best performer classifier interpretability (scripted task)*

If we were to remove the three variables (i.e., baseline IDS, F2 bandwidth variability and average pause duration) the confidence in classification would drop from 64% to 34% (0.64-0.12-0.1-0.08 = 0.34).

Looking at True Negatives, the classifier is 99% confident that this is a less severe case due to low values of F2 bandwidth variability, baseline IDS, mean F2 bandwidth, average syllable duration and high values for average pause duration and mean Pitch (F0).

Analysing the interpretability of the Feedforward Neural network (Figure 8) when deriving classification during the unscripted task, when considering True Positives, the model has a confidence of 67% in classifying this as a positive case. High values for mean F2 bandwidth, baseline IDS and for the fraction of unvoiced frames increase, increase the chance for an instance to be predicted as severe during unscripted speech tasks.

The results for the True negatives for the same cohort (Figure 8), reveal the model to be 79% confident that this is a less severe depression case. High values for Pitch (F0) mean, lower HNR values, lower baseline GAD7, higher average syllable duration contribute to the chance of being classed as a less severe case of depression.

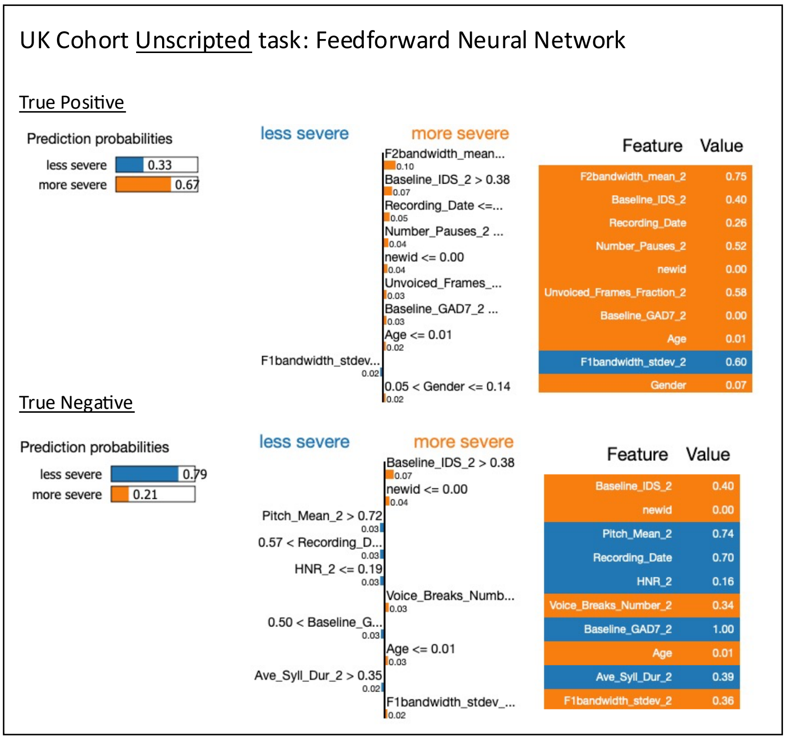

Figure 8. *UK cohort best performer classifier interpretability (unscripted task)*

6.4.5.2.Dutch cohort

The AUC for the Dutch cohort for either the best performer classifier in the scripted or unscripted task was lower than the one achieved for the UK cohorts in either model. Consequently, it is hypothesised, the confidence levels from the LIME algorithm when deriving interpretation especially for the True positives is average (Figures 9 and 10) due to this fact.

Starting with the classifier built for the scripted task (Figure 9) and looking specifically into the True Negatives of this model, as it has a 78% confidence in classifying instances as less severe, lower values of F2 bandwidth variability and of baseline IDS, higher values of mean F1 bandwidth and of articulation rate are associated with a higher chance to be classed as a less severe case. Of interest to note is the speaking rate value which suggest lower speaking rates increase the chance of being classed as less severe depression.

Looking into the same classifier but for the True positives, of note is high values for variability of F2 bandwidth, lower syllable duration values and higher number of pauses increase the chance of an instance to be classified as a more severe cases of depression.

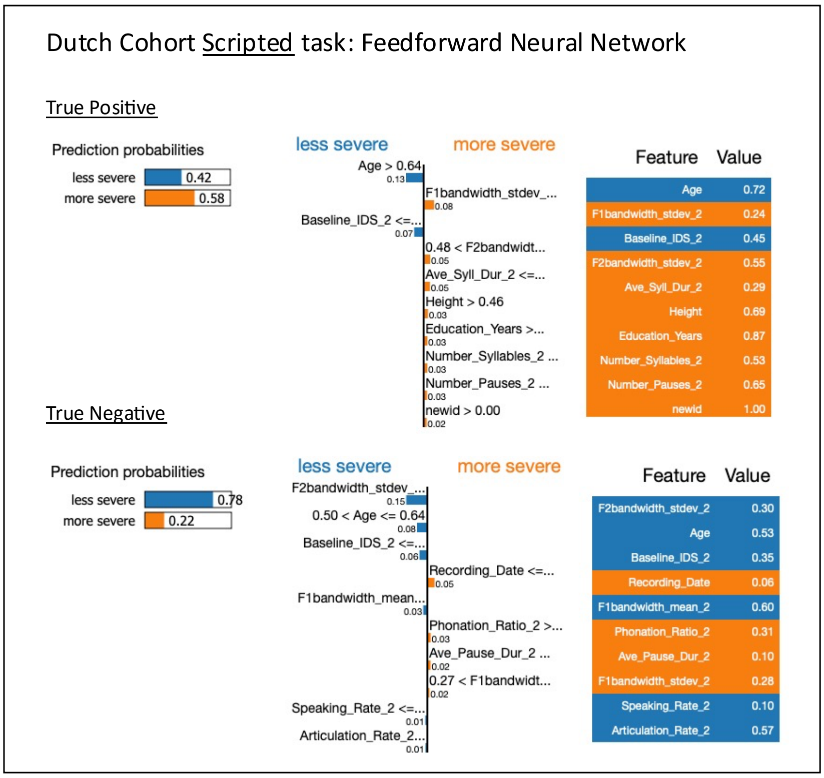

Figure 9. *Dutch cohort best performer classifier interpretability (scripted task)*

The Random Forest classifier for the unscripted task (Figure 10), when analysing the True Positives, revealed higher values of baseline IDS, mean F2 bandwidth, variability of F2 bandwidth, lower mean intensity (loudness) to be associated with higher severity of depression. Higher number of syllables is decreasing the classification as a more severe case. Overall, the results of the LIME algorithm for the unscripted task in classifying severity of depression are not very confident (51% for a true positive case and 56% for a True negative case)

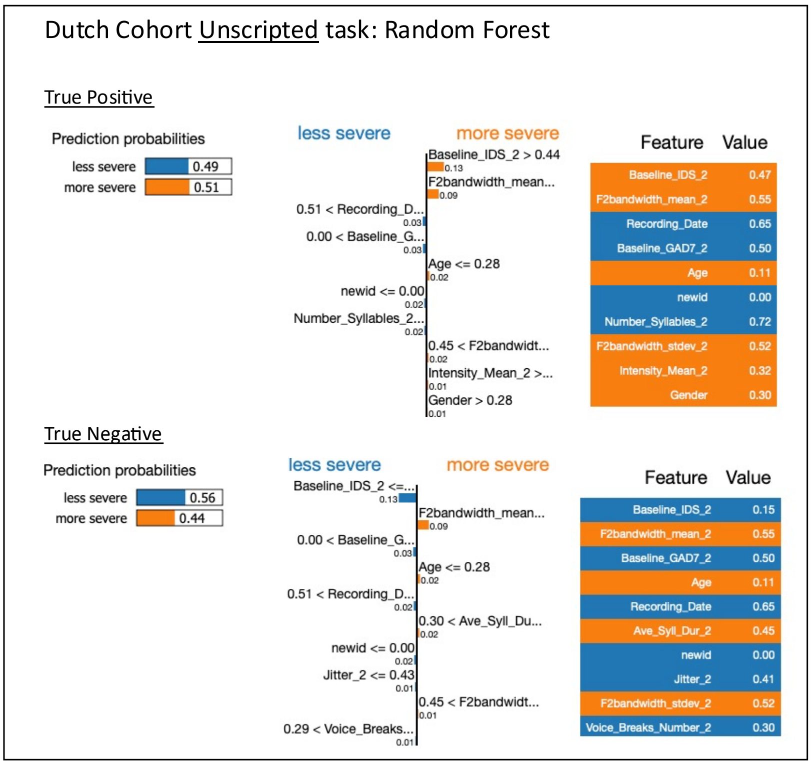

Figure 10. *Dutch cohort best performer classifier interpretability (Unscripted task)*

7. Discussion

We have observed in the results section that significant changes occurred for some of the speech features in the three different cohorts depending on depression severity. In this section the discussion will be organised according to the hypotheses we had before carrying out the linear mixed effects models as well as building the machine learning classifiers. The results will also be compared with the existing literature and theory discussed in section 3.

Hypothesis 1: Speech duration will be longer for depressed individuals. The results from the linear mixed effects models from the UK scripted task reveal an opposite relationship. There is a significance between speech/recording duration and depression but in the opposite direction. It was observed in the results section that for each one unit increase of recording time, there is a decrease in PHQ-8. Although this finding is contrary with the findings from

Mundt et al. (2007, 2012) it supports the findings from Albuquerque et al. (2021) where the authors also found depressed speech had a shorter duration. Although not significant for the UK non-scripted task, nor for the Spanish and Dutch cohorts, the evidence for these cohorts is that the trend is the same (Table 8). The results of the present project sustain that more severe depression speech reveals a shorter total speech duration.

Hypothesis 2: Longer pause times are associated with depression. Contrary to what was observed in the literature review (Alghowinem et al., 2013; Huang et al., 2019; Mundt et al., 2007, 2012; Stasak et al., 2019), when fitting the linear mixed effects model average pause duration was not significantly associated with severity of depression across any of the cohorts and speech tasks (Table 8 with detailed results). The author speculates because participants were recording themselves in the privacy of their homes, they could have repeated the recording and mitigate the pause time duration. However, this is highly unlikely due to the large number of speech recordings and the consistency of non-significance of the results of this feature across the three cohorts.

However, when looking into the machine learning classifiers interpretability through LIME, it was observed that higher values for average pause duration in the UK scripted task, were linked with increases of an instance being classed as less severe when classifying depression (Figure 7). The same was observed for the machine learning classifier built for the Dutch scripted task (Figure 9).

Still on the topic related with pauses, both in the UK and in the Netherlands during the scripted task, number of pauses was a speech feature found to be associated with severity of depression. These results reveal that an increase in pauses is associated with higher values of PHQ8 and therefore more severe depression.

For the Dutch cohort, on prosody measures related with pause during speech, it was observed pause frequency both in the scripted and unscripted speech tasks to be significantly associated with depression severity. This revealed pause frequency to have an inverse relationship with depression, the higher the pause frequency, the lower the severity of depression. Although not significant for the other two cohorts the trend was similar (Table 8 for complete details). To sum up, prosody measures related with pauses during speech revealed some interesting insights. Longer pause times in this project was not significantly associated with depression, number of pauses (UK and in the Netherlands during the scripted task) and pause frequency (Dutch cohort during the scripted and unscripted speech tasks) were significantly associated with depression.

Hypothesis 3: Slower speaking rates are associated with depression and are believed to be a good predictor for depression. Cannizzaro et al. (2004), Mundt et al. (2007, 2012) and Yamamoto et al. (2020) in their studies found a significance for the association between slower speaking rates and depression severity. In the present study, a significant association between speaking rate and depression severity was also found, however the relationship was different from the one in the hypothesis. For the Spanish cohort, for a one unit increase in speaking rate there is an increase in depression severity. This means that higher speaking rates in the case of the Spanish cohort are associated with more severe depression.

The linear mixed effects model analysis also revealed the average number of syllables between pauses (mean length run) to be higher in more severe cases of depression for the UK cohort during the scripted task. This finding together with the speaking rate significance, contradicts some of the speech science literature. The author speculates that this can be caused by language and cultural effects, i.e., in the case of Spanish cohort it may be the case that the language is causing the speaking rate effect, i.e., more depressed speech being faster for the Spanish language versus what has been observed for the English language in previous studies. This is an interesting hypothesis to be explored in future research. For the insight on the higher mean length run significance during the scripted task for the UK cohort, a possible explanation may be related with learning effects as participants got used to reading the poems during the 18-month duration of the study.

Hypothesis 4: Slower articulation rate is a speech feature that discriminates depressed and non-depressed individuals. This hypothesis is validated in the mixed effects models results. As per Table 8, for the Spanish cohort (scripted task) a significant relationship was observed between a slower articulation rate and severe levels of depression. It was seen that increases in articulation rate are associated with lower PHQ8 values, hence less severe depression.

When analysing the interpretability of the Dutch machine learning classifier for the scripted task (Figure 9) and specifically the True negatives explanations, it was also observed that higher values for articulation rate were linked with an instance being classed as less severe.

These insights confirm what has been observed in Albuquerque et al. (2021) study in the sense that slower articulation rates are associated with depression.

One other insight not necessarily related with articulation rate but is also a prosody speech feature associated with depression severity is the phonation ratio. This is linked with the phonation of sustained vowels, where it was possible to observe increases in the ratio of sustained vowels is associated with less severe cases of depression. In other words, lower values of phonation ratio, higher severity of depression.

Hypothesis 5: F0 is a speech feature able to discriminate depressed from non-depressed individuals. As previously seen in the literature review, mixed results were found for the association between F0 or Pitch and depression. The results found for the role of F0 in discriminating depression in this project confirm lower levels of F0 to be significantly associated with depression severity and confirm the results found in the literature review (Alghowinem et al., 2013; Darby et al., 1984; Low et al., 2020; Mundt et al., 2007). A significant association between F0 and severity of depression was found in this project in the

Dutch scripted and unscripted speech tasks. As per observed (Table 8), increase levels of mean F0 are associated with less severe cases of depression. This insight provides clarity on the debate seen to date for the role of F0. More specifically, for the Dutch cohort it validates that lower Pitch mean values (F0) are significantly associated with more severe depression. The results for the UK cohort, although not significant, reveal the same trend. This validates what has been found in the work of Alghowinem et al. (2013), Darby et al. (1984), Low et al. (2020) and Mundt et al. (2007) where lower values of F0 were associated with more severe depression.

Looking at the machine learning classifiers features interpretation, the classifier performing the best for the UK scripted cohort (Figure 7) also confirmed higher levels of Pitch mean to increase the chance of an instance to be classed as less severe for depression. The 99% confidence level for this model in interpreting an instance as true negative also reveals this insight. The same holds true for the insights of the classifier built for the UK unscripted task (Figure 8).

Hypothesis 6: Loudness will be lower for individuals with higher severity of depression. Like what is observed in the literature for mixed results for F0, loudness also presents different results for its association with depression in the literature (Cummins et al., 2015). This project confirms the initial hypothesis that loudness is lower for individuals with severe depression. In the UK scripted and unscripted speech tasks the association between loudness and severity of depression is significant. For each unit increase of loudness (intensity) there is a reduction in severity of depression. As a result, the hypothesis that loudness is lower for individuals with higher severity of depression is confirmed.

Similar insights could be derived when looking into the LIME interpretations for the Random Forest classifier built to classify depression severity using the unscripted task features (Figure 10). In this model one of the features contributing for a classification of more severe depression is a lower value of mean Intensity/loudness.

Hypothesis 7: Voice features derived from free/spontaneous speech and automated/read speech tasks offer different opportunities to discriminate severity of depression.

This hypothesis has been proved. Speech features associated with depression for the same cohort are different when explanations are being sought for the significance of the associations between speech features during scripted and unscripted tasks and depression, but also when looking from the point of view of classifying depression based on speech derived from scripted or unscripted tasks.

When looking from a statistical modelling perspective, it was possible to observe that speech features during scripted tasks are providing more statistically significant associations with depression severity than speech features whilst performing unscripted tasks. Taking the UK cohort as an example (Table 8), a total of 5 scripted features out of the 28 features analysed were significantly associated with depression severity, whilst only one speech feature from the unscripted task was associated with depression severity. The same trend was observed in the Spanish (4 speech features during the scripted task features were significantly associated with depression severity and only 2 during the unscripted task) and Dutch cohorts (6 speech features during the scripted task and 5 for the unscripted task).

In the case of using speech features during scripted and unscripted voice tasks for classification purposes (i.e., predicting severity of depression), the classifiers in the case of the UK cohort for example performed better with the features used during the scripted task.

These insights however do not support the findings from Alghowinem et al. (2013) in the sense that the authors observed, with the exception for shimmer and FO, speech features from spontaneous (unscripted) tasks produced better results in classifying depression. This was not observed in the present research. For this project it is concluded that speech from scripted tasks is providing better insights to discriminate severity of depression, both in exercises of statistical modelling to explain the relationships, and in exercises of building machine learning classifiers to predict severity of depression. However, this project used a much larger sample size than the one found in Alghowinem et al. (2013), which was limited to 30 depressed and control individuals, and is also unique on its multi-lingual and longitudinal nature as per discussed in section 2.

Hypothesis 8: Jitter is higher for depressed individuals and associated with depression. This hypothesis was validated, especially in the Dutch cohort where a significant association between jitter and severity of depression was found during the scripted task. Jitter was found to be higher for individuals with more severe depression. More specifically, for every unit increase of jitter, an increase in PHQ8 score is observed. This finding is in line with the results in Alghowinem et al. (2013) and Teixeira et al. (2013). Furthermore, looking into Figure 10 (LIME explanations for the Random Forest classifier for the unscripted task), it is possible to observe lower levels of jitter increase the chance of depression to be classified as less severe.

Hypothesis 9: shimmer is lower for depressed individuals and associated with depression. No statistically significant association was found between shimmer and depression for any of the cohorts and speech tasks. Also, in none of the explanations for the classifiers for the UK and Netherlands shimmer was given as one of the features contributing for a certain depression outcome. However as mentioned in the review done by Cummins et al. (2015) the lack of standardization in extracting segments of speech for shimmer analysis makes it difficult to produce reliable results. This may be a reason for the non-significance of shimmer in this project.

Hypothesis 10: HNR is lower for depressed individuals and associated with depression. Contrary to this study hypothesis, HNR was found to be higher for individuals with more severe depression (Table 8). Both in the UK cohort during scripted tasks and in the

Dutch cohort during scripted and unscripted tasks, higher HNR levels were associated with more severe depression. Despite significance not found for the UK unscripted task and for the Spanish cohort in both speech tasks, the relationship had the same trend as the one found for the cohorts with significance. This was also confirmed when looking at the explanations provided by LIME for the UK cohort during classification for the unscripted task. Lower values of HNR increase the chance of depression to be classified as less severe (Figure 8).

Hypothesis 11: Depressive speech is characterised by lower F1 and F2 versus normal speech. When looking at F1 measures, F1 frequency mean is significantly associated with depression severity in the Spanish cohort during scripted and unscripted speech tasks and with the Dutch cohort during the scripted task. These reveal somehow mixed results. In the case of the Spanish cohort, increases in Mean F1 frequency in both tasks are associated with lower severity of depression. In other words, lower levels are significantly associated with more severe depression for the Spanish cohort. Whilst in the case of the Dutch cohort, higher levels of Mean F1 frequency are associated with more severe depression, meaning lower levels of F1 frequency mean are linked with less severe depression. The author hypothesises that this may the case due to language differences.

Mean F1 bandwidth is also significantly associated with severity of the depression for the unscripted task in the Spanish cohort. Higher levels of Mean F1 bandwidth are associated with more severe depression. Lastly, variability in F1 bandwidth is also associated with severity of depression in the case of the Dutch cohort during unscripted voice tasks. This same insight (variability in F1 bandwidth) could also be observed through the LIME explanations (Figure 8) for the Feed forward neural network classifier explanations of the UK unscripted task. Higher levels of F1 bandwidth standard deviation (variability) are contributing for a depression classification to have the chance to be classified as less severe.

Linear mixed effects models did not find a significance between any of the F2 measures analysed and depression. However, it is of importance to note when looking at the interpretation of the machine learning classifiers for the UK cohort that variability in F2 bandwidth in the UK machine learning classifier for scripted tasks is contributing for depression to be classed as more severe when variability is high. Therefore, high values of F2 bandwidth variability seem to be contributing to classify cases as more severe in the case of the UK cohort scripted classifier (Figure 7). The same insight can be observed in Figure 9 for the Dutch cohort machine learning classifier interpretability for scripted tasks. Higher values of variability in F2 bandwidth are increase the chances of cases to be classified as more severe depression.

Higher values of Mean F2 bandwidth are also contributing for decisions to be classed as more severe depression as per observed in the explanations of the machine learning classifier for the UK cohort unscripted task (Figure 8).

Hypothesis 12: Prosody and voice acoustics features are good basis to classify depression patients across different geographies. This hypothesis has been proved. However, it is important to note that speech features alongside demographics (age, gender, height, and education years), baseline depression scores (baseline PHQ8, baseline IDS) and baseline anxiety scores (baseline GAD7) have proved to outperform in both the UK and Dutch cohorts the classifiers that only included the 28 speech features. This suggests that context is also important when it comes to classify depression. Yes, speech features are very important as per observed also in the interpretations provided by LIME as for the features contributing for the binary classification, however context is also of high importance when it comes to depression severity classification. When building machine learning classifiers to predict depression severity, the results from this project sustain that demographics, baseline depression and anxiety scores should be included alongside speech features as predictors.

It was possible to observe in section 6.4 (and in Table 9) that the classifiers built with the UK cohort data (both scripted and unscripted) had better AUC performance versus the Dutch classifiers. This could be associated in one hand to the fact that the Dutch cohort had a smaller sample size (UK cohort scripted task: 248 participants; Dutch cohort scripted task: 98) and in the other that the classifiers would benefit from additional hyperparameter tuning for the Dutch cohort to cater for language differences.

Hypothesis 13: Speech prosody and quality features changes are consistent across cultures. We have observed both from the linear mixed effects model results and from the machine learning classifiers LIME explanations that the speech features across cultures vary considerably. Therefore, this hypothesis given the results provided in this project is refuted. Language and culture do seem to play a role in voice prosody and acoustics features associated with depression. None of the speech features were significantly consistent across the three cohorts. With the exception for number of pauses, intensity mean, HNR, F1 frequency mean which were significantly in more than one cohort all the other significant speech features were very specific to the respective cohort.

Notwithstanding the promising results of this project, it is worth highlighting some of the limitations of the present study. As per seen in the results section, there is a clear skew towards females across the three cohorts. This has been mitigated by adding gender as a confounding variable in the linear mixed effects modelling and it has also been included in the machine learning classifiers as a predictor. The sample sizes of the Dutch and Spanish cohorts are lower than the UK one. To minimize this, the data was standardised to a mean of zero and a standard deviation of one to control for differences. The frequency of speech recordings both for the scripted and unscripted tasks was lower for the Spanish cohort, to mitigate this and to understand if the odds of providing speech samples were associated with depression severity, logistic regression models were run for the three cohorts as per seen in section 6 (detail of results in Table 7). Results, for the Spanish cohort, showed no significance was found associating severity of depression to the odds of providing more frequent data.

8. Conclusion

This project offered a unique opportunity to validate previously speech biomarkers found in controlled laboratory experiments and establish speech as a digital phenotype for depression. The study analysed, as far as the author is aware, the largest longitudinal and multilingual cohort study done to date on depression relapse using remote assessment technologies. Specific changes in prosody and voice acoustics features have been observed to significantly be associated with the onset of depression severity. In line with previous research findings, higher number of pauses revealed to be associated with the onset of more severe depression (in the UK and in the Netherlands cohorts during the scripted task), slower articulation rates were associated with more severe depression (in the Spanish scripted task), lower values of phonation ratio were associated with depression severity (Spanish scripted task). On measures of voice quality, pitch (F0) mean values were lower amongst more severe cases of depression (Dutch cohort both speech tasks), loudness was found to be lower amongst more severe cases of depression and Jitter was found to be higher for individuals with more severe depression (Dutch cohort scripted tasks).

This project also reported differences from results previously found in the literature. For example, speech duration revealed to be lower for individuals suffering from more severe depression (UK scripted task), higher speaking rates in the case of the Spanish cohort were associated with more severe depression, higher pause frequency in the Dutch cohort (both tasks) was found to be associated with less severe depression, HNR was found to be higher for individuals with more severe depression both in the UK (scripted task) and in the Dutch cohorts (both tasks). It was also concluded that speech from scripted tasks provided better insights to discriminate severity of depression contradicting previous results found in the literature (Alghowinem et al., 2013).

Some of the study hypotheses have been refuted. More precisely, longer pause times in the more depressed group was not statistically significant. Shimmer was also not statistically significant in the association with depression severity. Linear mixed effects models also revealed no significance between any of the F2 measures analysed and depression.

Speech was confirmed to be a promising biomarker to be used as an effective foundation to predict onset of severity depression. Together with other predictor variables, the results of the machine learning classifiers offered robust results. Machine Learning classifiers built using the UK cohort data as the foundation offered the best AUC results in the project. Classifiers using speech features from scripted speech tasks together with other predictors such as demographics, baseline depression and anxiety scores performing strongly.

The project results are encouraging in the sense that they establish a strong foundation for speech as a biomarker for depression severity diagnosis. Future research could further examine the language and cultural differences when it comes to speech features as biomarkers for depression.

References

Albuquerque, L., Valente, A. R. S., Teixeira, A., Figueiredo, D., Sa-Couto, P., & Oliveira, C. (2021). Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan. *PLOS ONE*, *16*(4), e0248842. https://doi.org/10.1371/journal.pone.0248842

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., & Parker, G. (2013). Detecting depression: A comparison between spontaneous and read speech. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7547–7551. https://doi.org/10.1109/ICASSP.2013.6639130

Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., & Snyder, P. J. (2004). Voice acoustical measurement of the severity of major depression. *Brain and Cognition*, *56*(1), 30–35. https://doi.org/10.1016/j.bandc.2004.05.003

Chawla, G. (2019). Feature request: Group aware Time-based cross validation. Github website. Accessed 1 August 2022, <https://github.com/scikit-learn/scikit-learn/issues/14257>

Cummins, N., Matcham, F., Klapper, J., & Schuller, B. (2020). Artificial intelligence to aid the detection of mood disorders. In *Artificial Intelligence in Precision Health* (pp. 231–255). Elsevier. https://doi.org/10.1016/B978-0-12-817133-2.00010-0

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, *71*, 10–49. https://doi.org/10.1016/j.specom.2015.03.004

Darby, J. K., Simmons, N., & Berger, P. A. (1984). Speech and voice parameters of depression: A pilot study. *Journal of Communication Disorders*, *17*(2), 75–85. https://doi.org/10.1016/0021-9924(84)90013-3

Doryab, A., Min, J. K., Wiese, J., Zimmerman, J., & Hong, J. I. (2014). *Detection of behavior change in people with depression*. 5.

Flint, A. J., Black, S. E., Campbell-Taylor, I., Gailey, G. F., & Levinton, C. (1993). Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. *Journal of Psychiatric Research*, *27*(3), 309–319. https://doi.org/10.1016/0022-3956(93)90041-Y

Horwitz, R., Quatieri, T. F., Helfer, B. S., Yu, B., Williamson, J. R., & Mundt, J. (2013). On the relative importance of vocal source, system, and prosody in human depression. *2013 IEEE International Conference on Body Sensor Networks*, 1–6. https://doi.org/10.1109/BSN.2013.6575522

Huang, Z., Epps, J., & Joachim, D. (2019). Investigation of Speech Landmark Patterns for Depression Detection. *IEEE Transactions on Affective Computing*, 1–1. https://doi.org/10.1109/TAFFC.2019.2944380

International Phonetic Association, & International Phonetic Association Staff (1999). Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge University Press.

Jeong, T., Klabjan, D., & Starren, Justin. (2016). *Predictive Analytics Using Smartphone Sensors for Depressive Episodes*. 7.

Kaggle (2020). Found the Holy Grail: GroupTimeSeriesSplit. Kaggle website. Accessed 1 August 2022, <https://www.kaggle.com/code/jorijnsmit/found-the-holy-grail-grouptimeseriessplit/notebook>

Kohn, R., Saxena, S., Levav, I., & Saraceno, B. (2004). The treatment gap in mental health care. *Bulletin of the World Health Organization*, 14.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population.

Journal of Affective Disorders, *114*(1–3), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, *5*(1), 96–116. https://doi.org/10.1002/lio2.354

Matcham, F., Barattieri di San Pietro, C., Bulgari, V., de Girolamo, G., Dobson, R., Eriksson, H., Folarin, A. A., Haro, J. M., Kerz, M., Lamers, F., Li, Q., Manyakov, N. V., Mohr, D. C., Myin-Germeys, I., Narayan, V., Bwjh, P., Ranjan, Y., Rashid, Z., Rintala, A., … Hotopf, M. (2019). Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): A multi-centre prospective cohort study protocol. *BMC Psychiatry*, *19*(1), 72. https://doi.org/10.1186/s12888-019-2049-z

Mental Health Foundation 2021. Mental Health Foundation website. Mental Health Foundation website. Accessed 1 August 2022, <https://www.mentalhealth.org.uk/our-work/research/coronavirus-mental-health-pandemic-study/wave-13-summary>

Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., & Geralts, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *Journal of Neurolinguistics*, *20*(1), 50–64. https://doi.org/10.1016/j.jneuroling.2006.04.001

Mundt, J. C., Vogel, A. P., Feltner, D. E., & Lenderking, W. R. (2012). Vocal Acoustic Biomarkers of Depression Severity and Treatment Response. *Biological Psychiatry*, *72*(7), 580–587. https://doi.org/10.1016/j.biopsych.2012.03.015

Radar-CNS (2016). Radar-CNS website, accessed 1 August 2022, <https://www.radar-cns.org>

Ribeiro, M.T.C. (2021). LIME, GitHub website. Accessed 1 August 2022, <https://github.com/marcotcr/lime>

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *'Why Should I Trust You?': Explaining the Predictions of Any Classifier* (arXiv:1602.04938). arXiv. http://arxiv.org/abs/1602.04938

Rush, A. J., Carmody, T., & Reimitz, P.-E. (2000). The Inventory of Depressive Symptomatology (IDS): Clinician (IDS-C) and Self-Report (IDS-SR) ratings of depressive symptoms. *International Journal of Methods in Psychiatric Research*, *9*(2), 45–59. https://doi.org/10.1002/mpr.79

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092. https://doi.org/10.1001/archinte.166.10.1092

Stasak, B., Epps, J., & Goecke, R. (2019). Automatic depression classification based on affective read sentences: Opportunities for text-dependent analysis. *Speech Communication*, *115*, 1–14. https://doi.org/10.1016/j.specom.2019.10.003

Stata (2022). Multilevel mixed-effects models, Stata website, accessed 1 August 20222, <https://www.stata.com/features/multilevel-mixed-effects-models/>

Steinhubl, S. R., Muse, E. D., & Topol, E. J. (2015). *The emerging field of mobile health*. 6.

Stevenson, D., Farmer, P. (2017). Thriving at Work: a review of mental health and employers. Gov.uk / Independent review. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/658145/thriving-at-work-stevenson-farmer-review.pdf

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters. *Procedia Technology*, *9*, 1112–1122. https://doi.org/10.1016/j.protcy.2013.12.124

Williams, T., Davis, J., Figueira, C., Vizard, T. (2021). Coronavirus and depression in adults, Great Britain: January to March 2021. Office for National Statistics. Accessed 1 July

2022,<https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/coronavirusanddepressioninadultsgreatbritain/januarytomarch2021>

World Health Organization 2022. *World Health Organization website*. Accessed 1 July 2022, < https://www.who.int/health-topics/mental-health#tab=tab_2>

XGBoost developers (2021). XGBoost Python API Reference. Accessed 1 August 2022, < https://xgboost.readthedocs.io/en/latest/python/python_api.html>

Yamamoto, M., Takamiya, A., Sawada, K., Yoshimura, M., Kitazawa, M., Liang, K., Fujita, T., Mimura, M., & Kishimoto, T. (2020). Using speech recognition technology to investigate the association between timing-related speech features and depression severity. *PLOS ONE*, *15*(9), e0238726. https://doi.org/10.1371/journal.pone.0238726

Yang, Y., Fairbairn, C., & Cohn, J. F. (2013). Detecting Depression Severity from Vocal Prosody. *IEEE Transactions on Affective Computing*, *4*(2), 142–150. https://doi.org/10.1109/T-AFFC.2012.38

Zhang, O. (2015). Winning Data Science Competitions, Slideshare.net website. Accessed 1 August 2022, < https://www.slideshare.net/ShangxuanZhang/winning-data-science-competitions-presented-by-owen-zhang>

Appendix I

Table 10.

Extracts of poems participants read in the scripted recordings

Passage 1: The North Wind and the Sun were disputing which was the stronger, when a traveller came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveller take his cloak off should be considered stronger than the other.
Passage 2: Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveller fold his cloak around him; and at last the North Wind gave up the attempt.
Passage 3: Then the Sun shone out warmly, and immediately the traveller took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two

Appendix II

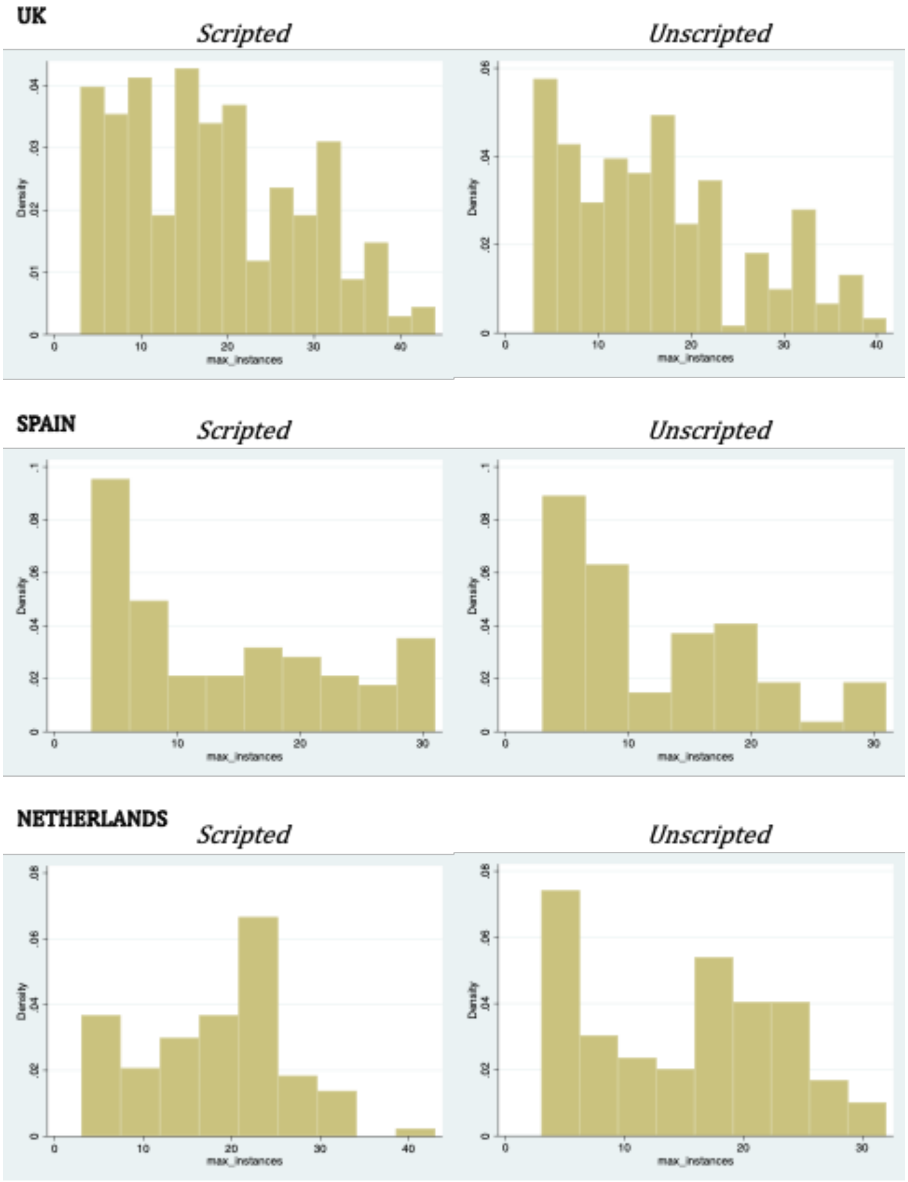

Figure 11. *Distribution of number of speech recordings provided per cohort and speech task*

Appendix III

Table 11.

Data imbalance depression binary classes (less severe depression and more severe depression): Training and Test sets

		Training sets		*Test sets*	
		Less severe	Severe	Less severe	Severe
UK	Scripted	59%	41%	53%	47%
	Unscripted	61%	39%	54%	46%
Netherlands	Scripted	58%	42%	61%	39%
	Unscripted	61%	39%	62%	38%