# Speech-Based Depressive Mood Detection in the Presence of Multiple Sclerosis: A Cross-Corpus and Cross-Lingual Study

**Monica Gonzalez-Machorro**[1,2,3], **Uwe Reichel**[1], **Pascal Hecker** [1,4], **Helly Hammer**[5], **Hesam Sagha**[1], **Florian Eyben**[1,6], **Robert Hoepner**[5], **Björn W. Schuller**[1,2,3,7]

[1]audEERING GmbH, [2]TUM University Hospital,
[3]Munich Center for Machine Learning, [4]Hasso-Plattner Institute,
[5]Inselspital, Bern University Hospital, [6]Agile Robots, [7]Imperial College
monica.gonzalez@tum.de

## Abstract

Depression commonly co-occurs with neurodegenerative disorders like Multiple Sclerosis (MS), yet the potential of speech-based Artificial Intelligence for detecting depression in such contexts remains unexplored. This study examines the transferability of speech-based depression detection methods to people with MS (pwMS) through cross-corpus and cross-lingual analysis using English data from the general population and German data from pwMS. Our approach implements supervised machine learning models using: 1) conventional speech and language features commonly used in the field, 2) emotional dimensions derived from a Speech Emotion Recognition (SER) model, and 3) exploratory speech feature analysis. Despite limited data, our models detect depressive mood in pwMS with moderate generalisability, achieving a 66% Unweighted Average Recall (UAR) on a binary task. Feature selection further improved performance, boosting UAR to 74%. Our findings also highlight the relevant role emotional changes have as an indicator of depressive mood in both the general population and within PwMS. This study provides an initial exploration into generalising speech-based depression detection, even in the presence of co-occurring conditions, such as neurodegenerative diseases.

## 1 Introduction

Depression is the most common psychiatric mood disorder (World Health Organization, 2023). Its prevalence is around 5% worldwide (World Health Organization, 2023). Despite its prevalence, depression often goes untreated (Johnson et al., 2022) due to factors such as socioeconomic barriers and a shortage of healthcare professionals (Evans-Lacko et al., 2018).

Speech-based Artificial Intelligence (AI) methods offer a promising approach for fast and non-invasive screening of neurological and mental health during routine examinations (Milling et al., 2022; Hecker et al., 2022), leveraging speech changes like reduced pitch, slower speaking rate, and articulation errors, which are common in individuals with depression (Cummins et al., 2015). These methods are accessible, scalable, and could enhance help-seeking behaviour and on-going monitoring (Johnson et al., 2022).

Prior work has utilised Machine Learning (ML) methods to detect depression using acoustic and linguistic features (Kappen et al., 2023). Mallol-Ragolta et al. (2019) trained a Recurrent Neural Network (RNN) on linguistic features for binary classification on the Distress Analysis Interview Corpus from the Wizard-of-Oz interviews (DAIC-WoZ) dataset, achieving an F1 score of 63%. Zhang et al. (2024) used wav2vec 2.0 for feature extraction and a Long Short-Term Memory (LSTM) network for binary classification using the DAIC-WoZ dataset, which yielded a 79% F1 score.

Similar work has also been conducted in other languages, such as for the German language, Menne et al. (2024) reported a balanced accuracy 88% for predicting depressive disorder against healthy controls using acoustic information, and for Italian language, in which Tao et al. (2023) reported an F1 score of 85% on the binary task of identifying depression using speech information from a reading task.

Automatic Speech Emotion Recognition (SER) research has also been effective in depression detection (Wang et al., 2020), for instance, Wang et al. (2021) developed a SER model on the DAIC-WoZ dataset for binary classification, reporting a 60% F1 score.

Depression is a common co-morbidity among people with neurodegenerative diseases, such as Multiple Sclerosis (MS), Parkinson's Disease (PD), and Alzheimer's Disease (AD), among others (Brenes, 2007), worsening both the Quality of Life (QoL) and disease prognosis (Hussain et al., 2020). In MS, for example, the lifetime risk of

depression is estimated around 50% (Arnett et al., 2008). The overlapping symptomatology of the two conditions can lead to misdiagnosis, with either one of them frequently overlooked (Hussain et al., 2020). While prior research highlights the potential of speech-based AI methods for depression detection (Cummins et al., 2015), further work is needed to assess their transferability in patients with neurodegenerative diseases like MS.

However, MS, due to its impact on the central nervous system, frequently leads to speech impairment, primarily dysarthria (Noffs et al., 2018). As a result, MS speech typically presents irregular articulatory breakdowns, distorted vowels, pitch breaks, harsh voice quality, and slow speaking rate (Noffs et al., 2018). This raises the question of whether speech-based depression detection can distinguish depressive symptoms in people with a co-existing speech impairment, such as dysarthria, due to a neurodegenerative disease, such as MS. We hypothesise that these methods would struggle to generalise and distinguish depressive symptoms in people with MS (pwMS), since some of the MS speech characteristics are similar to those found in people with depression.

This contribution aims to address this challenge by assessing the performance of common speech-based methods for depressive mood detection in pwMS. To do so, we conduct a cross-corpus and cross-lingual analysis using a well-known English-language corpus with depressive mood assessments, along with a German-language dataset of people with low MS disability, who also underwent depressive mood assessments. Our research questions are:

1. Do ML methods for depressive mood detection generalise to depressive mood detection in pwMS?

2. Given that SER models have shown promise in detecting emotional changes (Wang et al., 2021), which output from a fine-tuned SER model is more effective for depression detection: the model's final results (the classification or regression head output from a SER model) corresponding to the emotional dimensions –arousal, valence, and dominance– or the model's contextualised representations?

3. Can exploratory feature selection analysis improve generalisability of depression detection in pwMS?

This contribution is structured as follows. Section 2 introduces the datasets, features, and methods employed. Sections 3, 4, 5 present the results, limitations, and discussions. Finally, section 6 draws conclusions from the analysis.

## 2 Materials and Methods

### 2.1 Dataset

We employ two datasets: 1) The DAIC-WoZ depression dataset in English presented in (Gratch et al., 2014), and 2) a Swiss German dataset for pwMS collected under the scope of the COMMITMENT trial (Gonzalez-Machorro et al., 2023). The trial protocol was approved by national regulatory authorities and local ethic committee (BASEC-ID number 2021-02423) and registered on clinicaltrials.gov (NCT05561621). The DAIC-WoZ is a collection of semi-structured interviews containing speech samples of 189 participants (Gratch et al., 2014). It provides predefined speaker-independent training, development, and testing sets, and is segmented at the turn level (Valstar et al., 2016). The dataset includes scores from the Patient Health Questionnaire-8 (PHQ-8) self-assessed depression questionnaire.

The COMMITMENT (Prediction of Non-motor Symptoms in Fully Ambulatory MS Patients Using Vocal Biomarkers) dataset consists of 50 fully ambulatory pwMS and 20 control participants. Participants with MS have low levels of disability, with a median Expanded Disability Status Scale (EDSS) score of 1.0–indicating minimal impairment– and a min/max EDSS score of $0.0/3.0$, which indicates no disability to moderate disability but still walking unaided. For this paper, we only use the MS cohort. Details on the speech recordings are described in (Gonzalez-Machorro et al., 2023). Depressive mood scores for each participant are available using the Beck Depression Inventory-II (BDI-II) questionnaire. The dataset contains multiple speech tasks. However, in this paper, we utilise two spontaneous speech tasks from each patient: (1) describing the weather on the day of recording and (2) recalling a neutral memory prompted by the word "grass". These tasks are chosen because they elicit spontaneous speech and resemble the interview style of the DAIC-WoZ dataset. Data was collected using the AISoundLab web platform, which is a web app, in which each patient could navigate through a voice recording session under the supervision of a study nurse (Gonzalez-Machorro et al.,

2023). All participants provided informed consent prior to participation, and all data was pseudo-anonymised to protect patient privacy. The ethics consent unfortunately does not permit the publication of the recorded data.

In this paper, participants from the two datasets are categorised as having *depression* or *no depression* based on clinically validated threshold scores from two depression questionaries (BDI-II and PHQ-8). For the PHQ-8, participants with a score of 10 or higher are classified as having *depression* (Kroenke et al., 2001; Dhingra et al., 2011); and for the BDI-II participants with a score higher than 19 were defined as having *depression* (Beck et al., 1961). It is important to keep in mind that these scores serve as indicators of depressive symptoms rather than definitive clinical diagnoses of depression.

Audio files are downsampled to 16 kHz. Diarisation for the DAIC-WoZ data is performed using the turn-level segments provided for each speaker. A Voice Activity Recognition (VAD) algorithm[1] is applied to segment audio files from both datasets, which due to license restrictions, is not open-source. For consistency with previous work, we employ the same VAD parameter values as in (Gonzalez-Machorro et al., 2023). Transcripts are automatically obtained for each VAD segment using Whisper version 2 (Radford et al., 2023) with the *base* model for English and German language. For the DAIC-WoZ dataset, we merge the original training and development sets while the original testing set is left intact. The motivation is that due to the small dataset, we opt to use a Cross-Validation (CV) strategy for a more robust evaluation. The COMMITMENT dataset, as its purpose is purely for evaluating cross-corpus and cross-lingual generalisation, is not partitioned and it is used as an additional testing set.

Table 1 describes the metadata for both datasets across the different dataset partitions. Missing values for the questionnaires are dropped before processing. Models trained solely on the COMMITMENT dataset would likely over-fit due to insufficient participants with depressive symptoms to learn acoustic and linguistic markers of depression. Given the imbalance of the two classes, random oversampling with replacement for the two classes and a random seed of 42 is applied. To do so, we employ the package imbalanced-learn (Lemaître et al., 2017).

[1] provided by audEERING GmbH

Table 1: Metadata for the two datasets employed in this study and the train-test split.

| Subset | Dataset | Total Participants | Sex (F/M) | Depression / No Depression |
|---|---|---|---|---|
| Train | DAIC-WoZ | 135 | 59 / 76 | 42 / 93 |
| Test | DAIC-WoZ | 44 | 22 / 22 | 13 / 31 |
| | COMMITMENT | 50 | 37 / 13 | 4 / 46 |

### 2.2 Feature extraction

We extract six commonly used acoustic and linguistic feature sets, and normalise them per dataset using the Robust Scaler, which is robust against outliers. All features are extracted at a VAD segment-level.

1. The Wav2Vec2 contextualised representations of length 1024 correspond to the mean pooling of the encoder output. These representations are extracted using a publicly available fine-tuned Wav2Vec2 model for 3-dimensional SER task (Wagner et al., 2023).

2. SER-dimensions –arousal, valence, and dominance– are obtained using the same Wav2Vec2 SER model (Wagner et al., 2023). These features represent the final outputs of the model returned by the 2-layer multitask regression head (Wagner et al., 2023). By extracting both types of information –the contextualised representations and the emotion dimensions– from the Wav2Vec2 SER model, we aim to investigate which one is more effective for depression detection.

3. Praat features (Feinberg, 2022) are extracted using Nkululeko (Burkhardt et al., 2022) and correspond to 39 features, such as voice quality, shimmer, jitter, and duration. This type of features has shown significance for depression detection (Cummins et al., 2015).

4. extended Geneva minimalistic acoustic parameter set (eGeMAPS) (Eyben et al., 2016) is extracted using the Speech & Music Interpretation by Large-space Extraction (openSMILE) feature extraction tool (Eyben et al., 2010). It contains 22 acoustic features related to prosody, voice quality, and articulation. Previous work has reported promising results in

depression detection (Cummins et al., 2015). We employ the 88 functionals and summary statistics from these features.

5. The psycholinguistic feature set consists of 51 linguistic features that represent the syntactic complexity, the proportion of sentiment tokens, and the proportion of nouns, verbs, negations, adjectives, among others.

6. RoBERTa embeddings are extracted using a multilingual model –XLM Large RoBERTa (Conneau et al., 2020) –. These embeddings correspond to the $[CLS]$ pooling output applied to the last hidden states of the model. Each segment is defined with a maximum length of 512 tokens and represented by a size of 768.

### 2.3 Methods

We define the following three modelling scenarios to investigate whether ML methods for depressive mood detection generalise in the presence of MS:

**A) Baseline Performance:** Each feature set and model type is trained and evaluated on the DAIC-WoZ training and testing sets. This task establishes a baseline for model performance in depression detection.

**B) Generalisability Evaluation:** Each feature set and model type is evaluated on the DAIC-WoZ testing set –to ensure consistent performance on the general population– and the COMMITMENT dataset. The aim is to assess how well models trained on data from the general population (DAIC-WoZ) generalise to the pwMS data.

**C) Feature Selection Modelling:** Following an exploratory feature analysis on the DAIC-WoZ training set, the resulting significant features are used for training and evaluation. This task aims to improve model performance by selecting relevant features for depression detection. Two scenarios are investigated:

**C_A)** Models are trained and evaluated on the DAIC-WoZ training and testing sets using selected features. In other words, it is Task A with selected features. This task assesses whether feature selection improves performance within the general population.

**C_B)** Models are trained on the DAIC-WoZ training set using selected features and evaluated on both the DAIC-WoZ testing set and the COMMITMENT dataset. This scenario, equivalent to Task B with selected features, explores whether feature selection improves generalisability to pwMS data.

**Exploratory feature analysis.** To investigate which features are significant to distinguish between speakers with and without depression in the training set, we use the Mann-Whitney U test ($p < 0.05$) because it is non-parametric and does not require the assumption of a normal distribution. This makes it suitable for our data, where not all features follow a normal distribution. Additionally, it is more conservative than other statistical tests, reducing the risk of Type I errors. To quantify the effect size, we use Cohen-R (Cohen, 1988). Relevant features are found by selecting among the significant ones those with an $r \geq .30$. Corrections for Type 1 errors are not performed due to the large size of the feature sets, so that the aim of this analysis is restricted to explore acoustic and linguistic feature trends.

**Modelling.** We implement supervised ML classification for implementing the three modelling tasks. For reproducibility, we seed the pseudo-random number generation. The models used are Support Vector Machine (SVM), Random Forest (RF), and eXtreme Gradient Boosting (XGB). These supervised learning algorithms were selected due to their consistently strong performance across a wide range of classification tasks (Fernández-Delgado et al., 2014). Each model is trained using Grid search 5-fold speaker-independent CV on the training set.

The hyper-parameter values optimised for the Grid Search for each model are as follows: for SVM, $C \in [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10]$, the kernel options include `linear` and `rbf`, and the gamma parameter is chosen from `scale` and `auto`. For XGB, the number of estimators $\in [200, 300, 450, 500]$, the learning rate $\in [0.001, 0.01, 0.1, 0.2]$, the maximum tree depth $\in [4, 5, 6]$, the column subsample ratio $\in [1, 0.3, 0.5]$, and the subsample ratio $\in [0.8, 1]$. Lastly, for the RF model, the number of estimators $\in [50, 100, 300, 500, 800, 1000]$, the criterion is either `gini` or `entropy`, the minimum number of samples required to split an internal node is $\in [2, 3]$,

and bootstrap sampling is either `True` or `False`.

The optimal hyper-parameters identified through this process are then used to train the model on the entire training set. Class weights are calculated from the training set and are incorporated to address the class imbalance in the data.

**Evaluation.** We calculate speaker-level Unweighted Average Recall (UAR), F1-score, precision, and recall. Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) scores were also calculated at a speaker-level. Due to space limitations, only the ROC curves for the best-performing tasks are presented. We also compute the 95% Confidence Interval (CI) for the UAR. The CIs were calculated using 1000 bootstrapping iterations [2].

## 3 Results

### 3.1 Exploratory feature analysis

The Mann-Whitney U test is applied to each feature in the training set of the DAIC-WoZ dataset. Due to interpretability limitations, the Wav2Vec2 and the RoBERTa representations are excluded from the analysis. The number of significant ($p < 0.05$) features with a sufficiently high effect size ($r \geq 0.30$) identified per feature set are: 1) SER-dimensions: 1 feature–valence–; 2) Praat features: 33 out of 39 features; 3) eGeMAPS: 64 out of 88 functionals; 4) Psycholinguistic feature set: 18 out of 51 features. These selected features are used in the modelling task C_A and C_B to assess whether feature selection improves modelling performance. Figure 1 shows the valence distributions for the binary depression class ("no_depression" and "depression"), which is the only significant features found for the SER-dimensions.

### 3.2 Modelling Results

Table 2 shows UAR and its CIs, F1-score, precision, and recall for *depression* (Dep.) and *no depression* (No Dep.) classes, across the best-performing models and all feature sets. As we are tackling a binary classification problem, the chance-level UAR is 50%. The best result for Task A (Baseline Performance) with acoustic features is achieved using SVM and SER-dimensions (UAR: 73%), while the best result with linguistic features is achieved using SVM and RoBERTa embeddings (UAR: 56%). For

[2] https://github.com/luferrer/ConfidenceIntervals

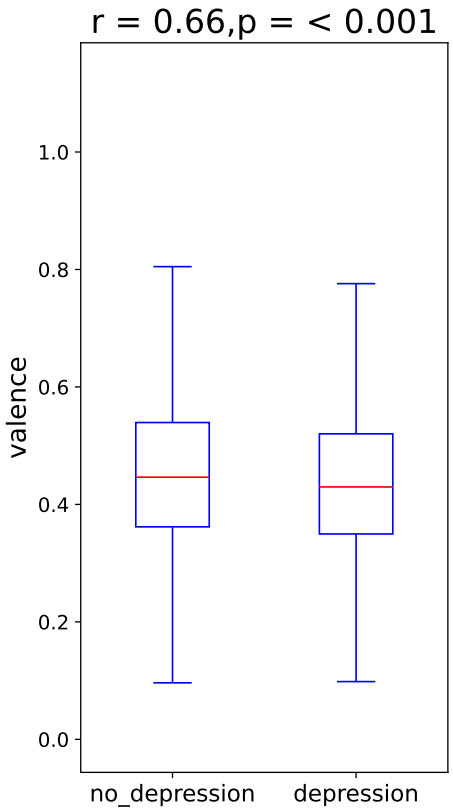


Figure 1: Feature distributions for the binary depression class depression class and valence dimension from SER in the DAIC-WoZ training set. This feature presents a $r$ of 0.66 and $p < 0.001$.

Task B (Generalisability Evaluation), Wav2Vec2 embeddings and Psycholinguistic features achieved the best performances (UAR: 66% and 62%, respectively). SER-dimensions in Task B show a performance drop. For Tasks C_A and C_B (Feature Selection Modelling on Tasks A and B), XGB with SER-dimensions obtained the highest UARs of 79% and 74%, respectively. Since SER-dimensions shows consistently good performance in all tasks, Figure 2 shows the ROC curves and AUC values for all tasks.

## 4 Discussion

In this paper, we explore three research questions: **1)** Do ML methods for depressive mood detection generalise to depressive mood detection in pwMS? Results in Table 2 indicate that for Task B (Generalisability Evaluation), acoustic-based features show reasonable generalisability to distinguish depression in pwMS, with only a modest performance decline compared to results from Task A (Baseline Performance).

In the case of the Wav2Vec2 features, a drop in

Table 2: Speaker-level test results. **A**: Baseline Performance. **B**: Generalisability Evaluation. **C_A**: Feature Selection on Task A. **C_B**: Feature Selection on Task B. The best-performing combinations for acoustic-based models are marked in **bold** and *; and linguistic models as **bold**$^{\dagger}$. Dep. corresponds to the *depression* class and No Dep. correspond to *no depression*

| Task | Feature | Model | UAR[%] | F1[%] | Precision[%] | | Recall[%] | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Dep. | No Dep. | Dep. | No Dep. |
| A | Wav2Vec2 | XGB | 66(49-81) | 65 | 81 | 47 | 71 | 62 |
| | SER-dimensions | SVM | **73(57-84)*** | **67*** | 48 | 90 | 85 | 61 |
| | Praat | XGB | 49(42-62) | 45 | 70 | 25 | 90 | 8 |
| | eGeMAPS | XGB | 54(46-69) | 53 | 72 | 50 | 94 | 15 |
| | Psycholinguistic | SVM | 46(32-63) | 46 | 68 | 25 | 61 | 31 |
| | RoBERTa | SVM | **56(48-71)**$^{\dagger}$ | **54**$^{\dagger}$ | 67 | 73 | 15 | 97 |
| B | Wav2Vec2 | SVM | **66(54-80)*** | **67*** | 50 | 88 | 41 | 91 |
| | SER-dimensions | SVM | 64(50-76) | 57 | 28 | 89 | 65 | 64 |
| | Praat | SVM | 47(33-60) | 39 | 16 | 79 | 53 | 40 |
| | eGeMAPS | XGB | 56(49-69) | 57 | 43 | 84 | 18 | 95 |
| | Psycholinguistic | SVM | **62(48-74)**$^{\dagger}$ | **54**$^{\dagger}$ | 26 | 88 | 65 | 60 |
| | RoBERTa | SVM | 55(49-67) | 54 | 50 | 83 | 12 | 97 |
| C_A | SER-dimensions | XGB | **79(70-87)*** | **70*** | 50 | 100 | 100 | 58 |
| | Praat | SVM | 51(36-68) | 51 | 31 | 71 | 31 | 71 |
| | eGeMAPS | XGB | 58(48-74) | 58 | 60 | 74 | 23 | 94 |
| | Psycholinguistic | SVM | 48(33-66) | 48 | 69 | 28 | 58 | 38 |
| C_B | SER-dimensions | XGB | **74(60-84)*** | **65*** | 37 | 93 | 76 | 71 |
| | Praat | SVM | 46(32-59) | 38 | 16 | 79 | 53 | 39 |
| | eGeMAPS | XGB | 57(46-70) | 56 | 29 | 84 | 29 | 84 |
| | Psycholinguistic | SVM | 54(42-68) | 51 | 22 | 84 | 41 | 68 |

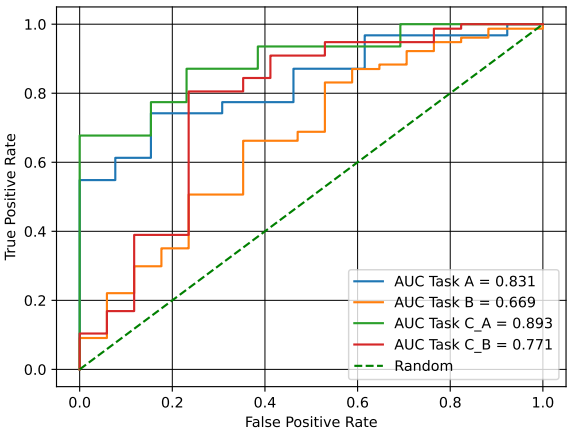


Figure 2: ROC curve and AUC value at a speaker-level for the best-performing models using the SER-dimenions as feature set across all tasks. Task A: Baseline Performance. B: Generalisability Evaluation. C_A: Feature Selection on Task A. C_B: Feature Selection on Task B.

performance for the two tasks is not found, which suggests that these features are transferable to other languages and groups with other co-morbidities such as MS. Interestingly, in the case of the eGeMAPS features, a minimal increase in performance is observed in Task B, which also suggests a generalisability capacity.

For Tasks C_A and C_B (Feature Selection Modelling), similar patterns are observed as in Tasks A and B, with SER-dimensions consistently outperforming other feature sets and demonstrating strong transferability in detecting depression among pwMS. This is further illustrated in Figure 2, which highlights the effectiveness of SER-dimensions in the context of MS.

The top-performing results for Tasks A (using SER-dimensions) and B (using Wav2Vec2 features) demonstrate greater precision in predicting the absence of depression (90% for "No Dep." in Task

A; 88% for "No Dep." in Task B) compared to predicting depression. This finding indicates that identifying depression using speech presents similar challenges in both same-language and cross-lingual contexts, as well as in the general population and among groups with co-morbidities, such as MS.

Interestingly, RF models did not outperform XGB or SVM in any task or feature set; consequently, they are excluded from Table 2. This was already reported by (Fernández-Delgado et al., 2014), where XGB has been shown to outperform RF in many cases.

**2)** Given that SER models have shown promise in detecting emotional changes, which output from a fine-tuned SER model is more effective for depression detection: the model's final predictions corresponding to the emotional dimensions or the model's contextualised representations? As shown in Table 2, the SER-dimensions and Wav2Vec2 representations achieve the highest UAR for Task A and Task B, respectively. SER-dimensions also outperform all other feature sets in Task C reaching the highest performance. Likely due to the high dimensionality of the Wav2Vec2 embeddings, SER-dimensions show overall better results by a small margin. However, the performance of SER-dimensions and Wav2Vec2 features heavily relies on the performance of the underlying SER model (Wagner et al., 2023), which was finetuned using the MSP-Podcast dataset (English language) (Lotfian and Busso, 2019). It is, therefore, unclear the cross-lingual generalisability of these features when training data would include languages other than English.

**3)** Can feature selection improve generalisability of depression detection in pwMS? Results for acoustic-feature-based models, with the exception of the Praat features, suggest that indeed, feature selection can improve the performance of depression detection. The feature analysis for SER-dimensions reveals that only valence among the three dimensions is significantly predictive, highlighting its important role as an indicator of depression in both the general population and pwMS. This finding is illustrated in Figure 2, which shows that individuals without depressive symptoms tend to use higher positive valence in spontaneous interviews compared to those with depressive symptoms. This aligns with prior research, such as (Trifu et al., 2024), which found that individuals with depression display lower positive valence than those without. This pattern may be attributed to a core symptom of depression: emotional dysfunction characterised by a predominant negative emotional state (Yang et al., 2023).

## 5 Limitations

In the case of text-based models, RoBERTa embeddings achieve above-chance performance in both Task A and Task B while psycholinguistic-feature-based models exhibit an unexpected trend: their performance on Task B surpassed that of Task A, C_A, and C_B. The suboptimal performances of text-based models may be due to the use of VAD segments for feature extraction, which ensured a consistent preprocessing pipeline across acoustic and text features, enabling direct comparisons between model types in detecting depression. While VAD segments effectively captured acoustic cues, contributing to strong performances, their short duration may have been less optimal for text-based features, such as word class proportions, which benefit from longer discourse contexts. The language-specific nature of these features also might have contributed to their struggle to generalise to the German-speaking MS population. Future work should explore longer segments to optimise text-based models, building on this study's foundation.

A limitation of this contribution arises from the use of different languages, recording conditions, and depression assessments. Although we try to tackle this by feature normalisation and the restriction to spontaneous speech, further research should explore the impact of language, depression assessments, and recording variations on the generalisability of speech-based depression detection. In this paper, we cannot definitively differentiate the extent to which the drop in model performance when evaluated on the MS population is influenced by language differences, recording conditions or the presence of MS itself.

Moreover, since both MS and depression are heterogenous conditions (Gaitán and Correale, 2019), implementing personalised approaches when screening for depression in pwMS is a crucial next step. Future work should also explore different stages of MS –this study focused on low-disability patients– and account for other co-morbidities in MS, like fatigue and cognitive decline, which may also influence speech. Also, the MS cohort was receiving pharmacological treatment, including com-

mon antidepressants for those MS patients diagnosed with depression, that could influence mood and, consequently, speech patterns. Although the general population diagnosed with depression from the DAIC-WoZ dataset may also have been undergoing pharmacological treatments, this information is not available in the dataset, preventing analysis of this potential confounding factor.

To further evaluate the transferability of speech-based depression detection, it is important to examine other common diseases where depression is a common co-morbidity and speech is impacted, such as PD or AD. A lack of depression scores in speech datasets for these disorders is a major limitation in this regard. Finally, acoustic and linguistic features alone cannot fully capture the multifaceted nature of depression. These ML methods are intended to augment established screening approaches. Incorporating other bio-signals, such as physiological data, could not only enhance performance but also provide a more comprehensive understanding of the disorder.

## 6 Conclusion

In this cross-corpus and cross-lingual study, we explore the efficacy of speech-based depressive mood detection in the presence of MS and across English and German languages. Our findings highlight the significance of emotional dimensions –arousal, valence, and dominance– in identifying depressive symptoms, not only in the general population but also within pwMS. Additionally, acoustic feature sets like eGeMAPS also demonstrate potential for generalisability in this context. However, further research is needed to establish robust conclusions. This study, despite its limitations, represents a step forward towards the integration and generalisability of speech-based depression detection methods. Non-invasive speech-based AI systems for depression detection hold the potential to improve the QoL for individuals with this disorder, even in the presence of other illnesses.

## Ethics Statement

This research was conducted in strict compliance with ethical standards. Data were analysed transparently to minimise bias and ensure robust accuracy. We extend our gratitude to all speakers who donated their data. This study was partially funded by Biogen. We thank Biogen's team for coordinating the financial support of the study.

## References

Peter A. Arnett, Fiona H. Barwich, and Joe E. Beeney. 2008. Depression in multiple sclerosis: Review and theoretical proposal. *Journal of the International Neuropsychological Society*, 14(5):691–724.

Aaron T. Beck, Clyde H. Ward, Myer Mendelson, John Mock, and John Erbaugh. 1961. An inventory for measuring depression. *Archives of General Psychiatry*, 4(6):561–571.

Gretchen A Brenes. 2007. Anxiety, depression, and quality of life in primary care patients. *Primary care companion to the Journal of clinical psychiatry*, 9(6):437–443.

Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben, and Björn Schuller. 2022. Nkululeko: A tool for rapid speaker characteristics detection. In *2022 Language Resources and Evaluation Conference, LREC 2022*, pages 1925–1932. European Language Resources Association (ELRA).

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Erlbaum, Hillsdale, NJ.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Subash S. Dhingra, Kurt Kroenke, Matthew M. Zack, Tara W. Strine, and Lina S. Balluz. 2011. PHQ-8 Days: A Measurement Option for DSM-5 Major Depressive Disorder (MDD) Severity. *Population Health Metrics*, 9:11.

Sara Evans-Lacko, Sergio Aguilar-Gaxiola, Ahmad Al-Hamzawi, Jordi Alonso, Corina Benjet, Ronny Bruffaerts, Wai Tat Chiu, Silvia Florescu, Giovanni de Girolamo, Oye Gureje, Josep Maria Haro, Yanling He, Chiyi Hu, Elie G. Karam, Norito Kawakami, Sing Lee, Crick Lund, Viviane Kovess-Masfety, Daphna Levinson, Fernando Navarro-Mateu, and Graham Thornicroft. 2018. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the who world mental health (wmh) surveys. *Psychological Medicine*, 48(9):1560–1571.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka,

Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. opensmile - the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia (ACM MM)*, pages 1459–1462, Florence, Italy. ACM.

David R Feinberg. 2022. Parselmouth praat scripts in python.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181.

María I. Gaitán and Jorge Correale. 2019. Multiple sclerosis misdiagnosis: A persistent problem to solve. *Frontiers in Neurology*, 10.

Monica Gonzalez-Machorro, Pascal Hecker, Uwe D. Reichel, Helly N. Hammer, Robert Hoepner, Lisa Pedrotti, Alisha Zmutt, Hesam Sagha, Johan van Beek, Florian Eyben, Dagmar M. Schuller, Björn W. Schuller, and Bert Arnrich. 2023. Towards Supporting an Early Diagnosis of Multiple Sclerosis using Vocal Features. In *Proc. INTERSPEECH 2023*, pages 1518–1522.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Pascal Hecker, Nico Steckhan, Florian Eyben, Björn W. Schuller, and Bert Arnrich. 2022. Voice analysis for neurological disorder recognition–a systematic review and perspective on emerging trends. *Front. Digit. Health*, 4:842301.

Madiha Hussain, Prabhat Kumar, Sara Khan, Domonick K Gordon, and Safeera Khan. 2020. Similarities between depression and neurodegenerative diseases: Pathophysiology, challenges in diagnosis and treatment options. *Cureus*, 12(11):e11613.

Jemimah A. Johnson, Prachi Sanghvi, and Seema Mehrotra. 2022. Technology-based interventions to improve help-seeking for mental health concerns: A systematic review. *Indian J. Psychol. Med.*, 44(4):332–340.

Mitchel Kappen, Marie-Anne. Vanderhasselt, and George M. Slavich. 2023. Speech as a promising biosignal in precision psychiatry. *Neuroscience Biobehavioral Reviews*, 148:105121.

Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. 2001. The phq-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Reza Lotfian and Carlos Busso. 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn W. Schuller. 2019. A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews. In *Proceedings of Interspeech 2019*.

Felix Menne, Felix Dörr, Julia Schräder, Johannes Tröger, Alexandra König, and Lisa Wagel. 2024. The voice of depression: speech features as biomarkers for major depressive disorder. *BMC Psychiatry*, 24:794.

Manuel Milling, Florian B. Pokorny, Katrin D. Bartl-Pokorny, and Björn W. Schuller. 2022. Is speech the new blood? recent progress in ai-based disease detection from audio in a nutshell. *Frontiers in Digital Health*, 4.

Gustavo Noffs, Thushara Perera, Scott C. Kolbe, Camille J. Shanahan, Frederique M.C. Boonstra, Andrew Evans, Helmut Butzkueven, Anneke van der Walt, and Adam P. Vogel. 2018. What speech can tell us: A systematic review of dysarthria characteristics in multiple sclerosis. *Autoimmunity Reviews*, 17(12):1202–1209.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Fuxiang Tao, Anna Esposito, and Alessandro Vinciarelli. 2023. The androids corpus: A new publicly available benchmark for speech based depression detection. In *Proc. Interspeech*, pages 4149–4153.

Raluca Nicoleta Trifu, Bogdan Nemeș, Dana Cristina Herta, Carolina Bodea-Hategan, Dorina Anca Talaș, and Horia Coman. 2024. Linguistic markers for major depressive disorder: a cross-sectional study using an automated procedure. *Frontiers in Psychology*, Volume 15 - 2024.

Michel Valstar, Maja Pantic, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, and Roddy Cowie. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In

*Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–13.

Hongbo Wang, Yu Liu, Xiaoxiao Zhen, and Xuyan Tu. 2021. Depression speech recognition with a three-dimensional convolutional network. *Frontiers in Human Neuroscience*, 15.

Xusheng Wang, Xing Chen, and Congjun Cao. 2020. Human emotion recognition by optimally fusing facial expression and speech feature. *Signal Processing: Image Communication*, 84:115831.

World Health Organization. 2023. Depression. https://www.who.int/news-room/fact-sheets/detail/depression. Accessed: 2025-03-10.

Chaoqing Yang, Xinying Zhang, Yuxuan Chen, Yunge Li, Shu Yu, Bingmei Zhao, Tao Wang, Lizhu Luo, and Shan Gao. 2023. Emotion-dependent language featuring depression. *Journal of Behavior Therapy and Experimental Psychiatry*, 81:101883.

Xu Zhang, Xiangcheng Zhang, Weisi Chen, Chenlong Li, and Chengyuan Yu. 2024. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. *Scientific Reports*, 14:9543.