applied sciences

MDPI

Article

Speech-Based Depression Recognition in Hikikomori Patients Undergoing Cognitive Behavioral Therapy

Samara Soares Leal [1,*], Stavros Ntalampiras [1], Maria Gloria Rossetti [2], Antonio Trabacca [3], Marcella Bellani [2] and Roberto Sassi [1]

[1] Department of Computer Science, University of Milan, 20133 Milan, Italy; stavros.ntalampiras@unimi.it (S.N.); roberto.sassi@unimi.it (R.S.)
[2] Department of Neurosciences, Biomedicine, and Movement Sciences, University of Verona, 37134 Verona, Italy; mariagloria.rossetti@univr.it (M.G.R.); marcella.bellani@univr.it (M.B.)
[3] Scientific Institute IRCCS "E. Medea", Scientific Direction, 23842 Bosisio Parini, Italy; antonio.trabacca@lanostrafamiglia.it
* Correspondence: samara.soares@unimi.it

Abstract

Major depressive disorder (MDD) affects approximately 4.4% of the global population. Its prevalence is increasing among adolescents and has led to the psychosocial condition known as hikikomori. MDD is typically assessed by self-report questionnaires, which, although informative, are subject to evaluator bias and subjectivity. To address these limitations, recent studies have explored machine learning (ML) for automated MDD detection. Among the input data used, speech signals stand out due to their low cost and minimal intrusiveness. However, many speech-based approaches lack integration with cognitive behavioral therapy (CBT) and adherence to evidence-based, patient-centered care—often aiming to replace rather than support clinical monitoring. In this context, we propose ML models to assess MDD in hikikomori patients using speech data from a real-world clinical trial. The trial is conducted in Italy, supervised by physicians, and comprises an eight-session CBT plan that is clinical evidence-based and follows patient-centered practices. Patients' speech is recorded during therapy, and the Mel-Frequency Cepstral Coefficients (MFCCs) and wav2vec 2.0 embedding are extracted to train the models. The results show that the Multi-Layer Perceptron (MLP) predicted depression outcomes with a Root Mean Squared Error (RMSE) of 0.064 using only MFCCs from the first session, suggesting that early-session speech may be valuable for outcome prediction. When considering the entire CBT treatment (i.e., all sessions), the MLP achieved an RMSE of 0.063 using MFCCs and a lower RMSE of 0.057 with wav2vec 2.0, indicating approximately a 9.5% performance improvement. To aid the interpretability of the treatment outcomes, a binary task was conducted, where Logistic Regression (LR) achieved 70% recall in predicting depression improvement among young adults using wav2vec 2.0. These findings position speech as a valuable predictive tool in clinical informatics, potentially supporting clinicians in anticipating treatment response.

Keywords: speech depression recognition; machine learning; wav2vec2

Academic Editor: Douglas O'Shaughnessy

Received: 29 September 2025
Revised: 23 October 2025
Accepted: 31 October 2025
Published: 4 November 2025

Citation: Leal, S.S.; Ntalampiras, S.; Rossetti, M.G.; Trabacca, A.; Bellani, M.; Sassi, R. Speech-Based Depression Recognition in Hikikomori Patients Undergoing Cognitive Behavioral Therapy. *Appl. Sci.* **2025**, *15*, 11750. https://doi.org/10.3390/app152111750

1. Introduction

Major depressive disorder (MDD), also known as depression, is a potentially life-threatening mental illness that severely limits psychosocial functioning [1]. This disorder

represents a global health concern, with an estimated 4.4% of the world population affected, as reported by the World Health Organization (WHO) [2]. In Italy, for example, 9.85 million people were affected by mental health disorders in 2019, accounting for 16.6% of the population, with 5.3% diagnosed with depression [3]. After the COVID-19 pandemic, this number increased, with depression affecting about 6.4% of the population in 2022 [4]. The incidence of depression among youths is increasing, largely due to their predominantly indoor lifestyle and addiction to mobile devices [2]. This has led to the advent of the psychosocial condition hikikomori, which is typified by prolonged social isolation (SI) [5].

The diagnosis of MDD is typically made on the basis of assessment questionnaires administered by medical professionals, such as the Hamilton Depression Rating Scale (HDRS), or self-assessment questionnaires, including the Children's Depression Inventory (CDI) [6,7]. While these tools offer valuable insights, concerns remain regarding bias in professional analysis and the impact of subjective responses [8]. A number of treatment options are available for depression, including pharmaceutical and non-pharmaceutical [9]. However, the burden of depression is mainly due to the delay in diagnosis and the lack of effectiveness of treatments.

Recently, several studies have explored machine learning (ML) as a strategy for the early diagnosis of depression [10], given its greater flexibility and scalability compared to traditional biostatistical methods, enabling applications across diverse healthcare tasks [11]. Among the data types employed for ML models in the context of MDD diagnosis (video, electroencephalogram, and text), speech signals are particularly noteworthy from an applicative point of view [12]. This is due to their minimal intrusion and cost, as well as their convenient adaptability across a range of platforms [13]. Furthermore, speech directly expresses emotions and reflects neural modulation through motor and vocal acoustic changes, rendering it challenging to conceal symptoms of MDD [14].

Despite the advantages of speech-based depression recognition (SDR), the field still faces several challenges. A major limitation is the reliance on small and imbalanced datasets for model training [15,16]. In addition, the reliability of current MDD diagnostic applications remains uncertain, as highlighted by Martinengo et al. (2021) [17]; few applications offer self-guided cognitive behavioral therapy (CBT); and many raise privacy concerns due to the sharing of user data with third parties. Furthermore, most of them lack a patient-centered design, do not follow evidence-based clinical guidelines, and fail to complement the ongoing therapeutic relationship between patients and clinicians, with some even attempting to replace professional monitoring altogether.

In light of the above, this work explores the use of ML to predict depression treatment response in hikikomori patients from speech data collected in a real-world clinical trial, contributing to ongoing advances in AI for precision and predictive health. The clinical trial was conducted by clinicians from the Unit of Psichiatry, Azienda Ospedaliera Universitaria, Verona and the La Nostra Famiglia association in Puglia, Italy. The study case comprises pre- and post-treatment depression assessments ($T0$ and $T1$), along with an intervention consisting of eight cognitive behavioral therapy (CBT) sessions and ten hours of cognitive restructuring (CR), grounded in evidence-based, patient-centered practices [18].

The study cohort includes 35 patients diagnosed with hikikomori (24 young adults and 11 adolescents). Speech data were recorded during CBT sessions; however, due to privacy constraints and the ethical standards governing the trial, only acoustic parameters were processed. Specifically, no lexical content or personally identifiable information was used. Instead, Mel-Frequency Cepstral Coefficients (MFCCs) and wav2vec 2.0 embedding were extracted from the raw audio signals, and employed to train ML models aimed at predicting patient improvement over the course of CBT.

MFCCs are widely used in speech processing as handcrafted features that capture short-term spectral characteristics of speech based on the perceptual scales of human hearing [19]. In contrast, Wav2vec2.0 is a recent self-supervised learning approach that uses deep Transformer architectures to learn high-level, contextualized representations directly from raw waveforms. While MFCCs provide interpretable, low-level descriptors, Wav2vec2.0 embeddings capture more abstract, task-adaptive representations [20]. In this study, we compare these two approaches to evaluate whether modern deep learning embeddings outperform or complement traditional handcrafted features when modeling speech for depression recognition.

The novelty of this study lies in its demonstration of clinical integration rather than algorithmic innovation, emphasizing the translational value of the proposed framework for medical practice. First, model predictions are aligned with established HDRS/CDI thresholds (e.g., $\geq$50% reduction or absolute cutoffs for response and remission), thereby mapping outputs onto clinically meaningful categories such as responder versus non-responder. Second, the framework leverages up to eight therapy sessions per patient to generate both cross-sectional estimates of symptom severity and longitudinal trajectories of change (Y_{Δ}). This enables clinicians to monitor whether patients improve consistently or deteriorate, thus supporting treatment planning and timely intervention.

The paper is organized as follows: Section 2 reviews related work on ML approaches for SDR. Section 3 describes the proposed methodology. Section 4 outlines the experimental design, while Section 5 presents the results. Section 6 discusses and interprets the findings. Finally, Section 7 concludes the paper and outlines directions for future research.

2. Related Work

According to recent comprehensive reviews on SDR, including Leal et al. (2024) [16], Wu et al. (2023) [10], Low et al. (2020) [14], and Cummins et al. (2015) [21], key limitations continue to hinder clinical adoption. In particular, SDR systems often lack integration with CBT-based interventions and remain insufficiently validated within patient-centered, evidence-based settings.

This study addresses these gaps through a clinically grounded design that models depression evolution across CBT sessions, rather than focusing solely on static depression detection. Our methodology incorporates pre- and post-treatment assessments to align predictions with therapeutic progression. While based on established features like MFCCs, it also integrates modern embeddings (wav2vec 2.0) within a real clinical protocol.

Given the CBT-based treatment pipeline in our dataset, direct comparisons with prior SDR work are challenging. Nonetheless, several related studies using similar acoustic features and modeling approaches offer valuable reference points. The work of [12], for example, proposes deep learning architectures for SDR and reports an average accuracy of 97%. However, the speech samples were self-recorded and submitted via WhatsApp, raising concerns about audio provenance and consistency under clinical conditions.

In the study of [19], jitter, MFCCs, their derivatives, and spectral centroid were used as features for SDR, applying a Sequential Minimal Optimization (SMO) algorithm. Although the approach yielded promising results—achieving 85% accuracy—it raises concerns regarding the speech data collection method. A similar issue is observed in [22], where a dataset for SDR was created using 16 h of smartphone-recorded free speech; the Logistic Regression model achieved 72.9% accuracy in diagnosing MDD. In both studies, the use of spontaneous speech, without integration into a physician supervised CBT plan, poses challenges for clinical validation and alignment with evidence-based practices.

A previous study by [15] proposed an SDR approach using speech recorded during routine psychiatric evaluations and in uncontrolled environments for healthy controls.

While the Random Forest model achieved high accuracy (87.5%), the absence of a structured CBT protocol and dataset imbalances (e.g., age, gender, and admission criteria) limit the clinical generalizability and interpretability of the results.

The work of [20] fine-tuned wav2vec 2.0 embeddings with a 1D-CNN and an attention-based LSTM for binary depression classification, achieving an F1-score of 79% on the Distress Analysis Interview Corpus with Wizard-of-Oz (DAIC-WOZ), highlighting the effectiveness of pre-trained acoustic embeddings in low-resource settings. Ref. [23] also leveraged wav2vec 2.0 with DAIC-WOZ, reporting an accuracy of 96.5% and an Root Mean Squared Error (RMSE) of 0.1875.

Although both studies demonstrate the potential of wav2vec 2.0 for SDR, it is important to note that DAIC-WOZ consists of scripted, computer agent-led interviews, rather than real therapy sessions with clinicians. In contrast, our dataset involves speech recorded during clinician-supervised CBT sessions, with outcomes measured through pre- and post-treatment assessments. These differences make direct comparison with these works difficult, as our task involves modeling therapeutic progression over time—not static depression detection. Moreover, while DAIC-WOZ-based studies use binary labels (depressed vs. non-depressed), our target is to predict whether a patient improved (better or worse) after treatment. Our emphasis shifts from screening to outcome forecasting, providing a directly actionable comprehension of the subject's status in clinical practice.

3. Methodology

The following subsections describe the methodology adopted in this study, detailing the speech processing steps and ML models used throughout the experiments. We employed a set of models with heterogeneous learning paradigms to evaluate MFCCs and wav2vec2.0 embeddings [16]. Random Forest (RF) and XGBoost (XGB) were included as tree-based ensemble methods that have shown strong performance in tabular data. Support Vector Regression (SVR) was selected as a kernel-based approach that can model non-linear relationships. A feed-forward Multi-Layer Perceptron (MLP) was also applied to capture more complex feature interactions.

3.1. Data Collection

The process begins with the collection and secure storage of audio recordings from patients' CBT sessions conducted by the clinical team. Speech recordings were collected through the COD20 telemedicine platform developed by the University of Milan, which provides secure video consultation between patients and clinicians. Audio was captured using the AWS Chime SDK via the client-chime-sdk-media-pipelines library. The resulting files followed a standardized format as follows: AAC-LC codec, 48 kHz sampling rate, mono channel, and MP4 container. This infrastructure ensured consistent and secure audio capture across all therapy sessions.

The clinical trial selection criteria were a score higher than 42 on the Post Intervention 25-item hikikomori scale (HQ-25), indicating a severe social isolation [5]. A total of 62 patients have been recruited thus far, and they provided informed consent to participate in the trial. Of these, 35 completed all eight CBT sessions and responded to both the pre- and post-treatment depression assessments ($T0$ and $T1$), and they were therefore included in this study. This cohort consists of 57.14% female and 42.86% male participants and was divided into two groups as follows: 24 young adults and 11 adolescents, as shown in Figures 1 and 2. The final sample included 276 h and 40 min of speech recordings, comprising 199,200 segments of 5 s. Patient-attributed speech accounted for 47.3% of the total recorded audio (130.9 h). Each segment represents a continuous chunk of speech extracted from speech session recordings. The 5 s duration was chosen to strike a balance

between temporal resolution and feature stability. It is short enough to capture meaningful vocal cues yet long enough for reliable acoustic extraction. This approach is inspired by the thin slices theory [24], which posits that it is possible to assess psychological states, such as depression, through short audio segments. Future extensions will include robustness checks with alternative segment durations.

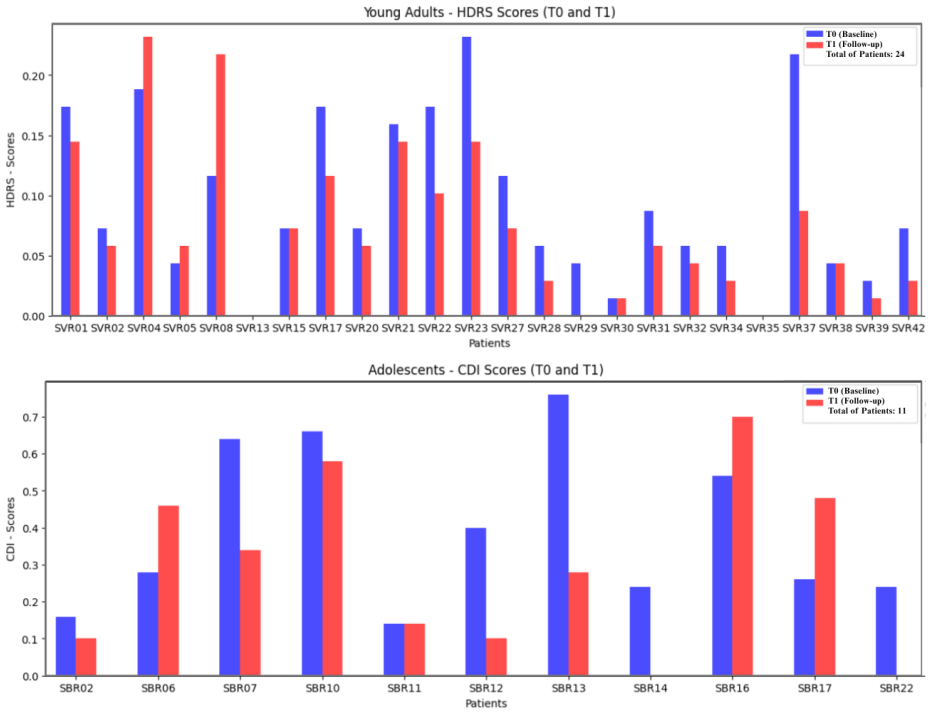

Figure 1. The distribution of patients across the groups.

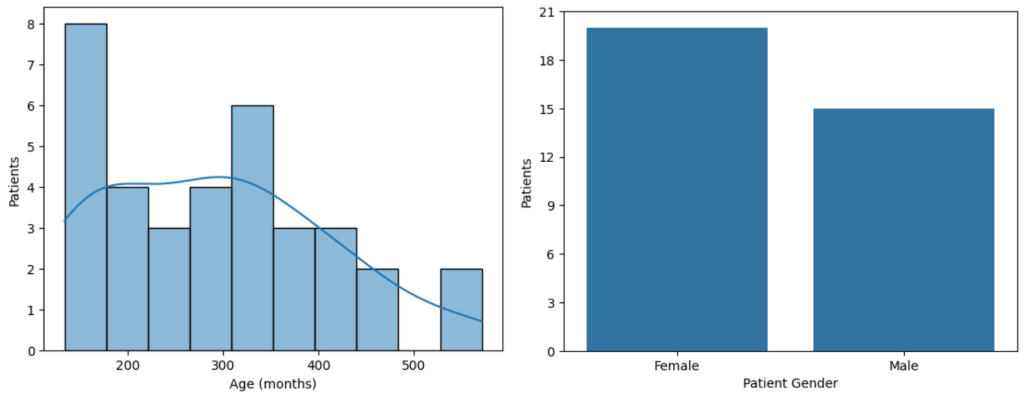

Figure 2. The distribution of age and gender across the population.

Depression in young adults (aged 18 and above) was assessed using the Hamilton Depression Rating Scale (HDRS), administered by medical professionals, with scores ranging from 0 to 69 [6]. In contrast, depression in the adolescent group (aged 13–17) was evaluated using the Children's Depression Inventory (CDI), a self-report questionnaire, with a score range from 40 to 90 [7].

Unfortunately, datasets related to mental health remain limited in size, primarily due to stigma and privacy-related concerns. For instance, the widely used DAIC-WOZ corpus includes 42 participants diagnosed with depression, while the Multi-modal Open Dataset for Mental-disorder Analysis (MODMA) comprises just 23 such cases [25]. Moreover, neither of these datasets includes speech recordings collected during clinician-supervised CBT sessions. Therefore, despite its relatively small size, the dataset presented in this study

represents a valuable contribution to the field, particularly due to its real-world clinical context. As an additional contribution, we are open to sharing the extracted speech-derived parameters (non-identifiable features) upon request to the corresponding author, in support of future reproducibility and research.

3.2. Preprocessing

In this step, to ensure that only the patient's speech was analyzed, a semi-automated identification procedure was applied to each session's audio recording. Clinicians first annotated short reference segments corresponding to the patient's speech. These annotations served as anchors for a k-means clustering algorithm ($k = 2$), which was then used to automatically assign the remaining audio frames to either the patient or another speaker. Only features extracted from segments identified as patient speech were retained for analysis. This process excluded therapist speech and background noise, improving the reliability of the extracted features.

3.3. Feature Extraction

Each therapy session's audio recording was divided into multiple 5 s segments. Each segment was treated as an independent data point for model training, enabling the fine-grained modeling of speech dynamics across sessions. From each segment, the mean and standard deviation were computed to construct the input feature vectors. Two types of acoustic features were extracted as follows:

1. **MFCCs**: For each segment, 13 MFCCs coefficients were computed using a 32 kHz sampling rate, 960-sample frames, and 75% overlap. The mean and standard deviation of each coefficient were then calculated within the segment. MFCCs capture the short-term power spectrum of speech on a Mel scale, reflecting perceptual aspects of human hearing [26,27].
2. **Wav2vec 2.0 embeddings**: A transformer-based, pre-trained self-supervised model for speech representation learning [28] was used to extract high-dimensional embeddings from each segment using a 16 kHz sampling rate. The mean and standard deviation were computed across all embedding dimensions.

3.4. Machine Learning Models

Machine learning (ML) refers to algorithms that learn patterns from data to make predictions [16]. In this study, we employed the models shown in Table 1. These models were chosen due to their ability to capture non-linear patterns and their compatibility with structured features such as acoustic and clinical variables, as well as their proven effectiveness in clinical prediction tasks.

Table 1. Overview of the models used in this study.

Model	Type	Main Characteristics
RF	Ensemble (Bagging)	Multiple decision trees; robust to noise.
XGB	Ensemble (Boosting)	Gradient boosting; strong regularization; efficient on tabular data.
SVR	Kernel-based	Effective in high-dimensional spaces; sensitive to kernel and parameters.
LR	Linear Model	Interpretable baseline; assumes linear separation.
MLP	Neural Network	Captures non-linearities.

The models were trained using a Leave One Patient Out (LOPO) protocol, in which data are split at the patient level as follows: in each fold, one patient is held out for testing while all others are used for training. This procedure maximizes data usage, prevents

information leakage across segments from the same individual, and provides a robust, low-bias estimate of the performance [29].

3.5. Output

The final output variable, Y_Δ, is defined as the standardized delta in depression scores ($T1 - T0$). To allow for joint modeling and to make prediction targets comparable, we applied min–max normalization to scale both scores to the $[0, 1]$ range, as detailed in Equations (1) and (2) and illustrated in Figure 2. This approach is commonly used in clinical data mining studies involving mental health conditions, particularly depression, as it allows models to learn from standardized patterns across different sources and populations [30]. This transformation ensures the comparability of prediction targets regardless of the underlying clinical scale. It preserves the relative severity within each scale while allowing the model to learn from the full cohort. Separate analyses were also conducted for each group to capture potential group-specific effects.

Each score was first min–max normalized to the $[0, 1]$ range as follows:

$$\tilde{T}^{HDRS} = \frac{T}{MAX\_HDRS}, \quad \tilde{T}^{CDI} = \frac{T - 40}{MAX\_CDI - 40}. \tag{1}$$

The standardized change was then calculated as

$$\Delta Y = \tilde{T}_1 - \tilde{T}_0, \tag{2}$$

which results in $\Delta Y \in [-1, 1]$.

For the classification task, thresholds were computed from outcome distributions and refined in consultation with the clinical team to ensure alignment with meaningful distinctions in depression severity. Following this guidance, the 30th and 70th percentiles of the Y_Δ distribution were adopted, resulting in approximately 34% of the samples being labeled as Worse, 34% as Stable, and 32% as Better. For the binary task, the 40th and 60th percentiles were used to define the class boundaries, yielding a distribution of approximately 55% for the Worse class and 45% for the Better class. Percentile-based thresholds were adopted to map normalized HDRS and CDI scores into the clinically meaningful categories of improvement, stability, and worsening. Since HDRS and CDI use different ranges and are not directly comparable, no unified cutoff can be applied across both populations; our approach ensured consistent class definition and clinical interpretability.

4. Experimental Design

The experiments were organized by groups as follows: young adults and adolescents, young adults only, and adolescents only. The following three prediction tasks were implemented for each group:

1. **Regression**: In this task, the standardized Y_Δ was used as the output variable. To contextualize model performance, two baseline experiments were conducted and evaluated using the same LOO-CV protocol as the ML models. Regression experiments were performed using RF, XGB, SVR, and MLP models.
2. **Binary**: This task aimed at enhancing the interpretability of depression treatment prediction using the RF, LR, XGB, and MLP models. Class labels were assigned based on a thresholding strategy (Better and Worse).
3. **Multiclass**: The multiclass task was also conducted using the RF, LR, XGB, and MLP models. In this approach, the class labels were Better, Stable, and Worse.

The following input settings were considered:

- **Setting 1:** MFCCs from the first session only.

- **Setting 2:** MFCCs from the first and the eighth session.
- **Setting 3:** MFCCs from all 8 sessions.
- **Setting 4:** Wav2vec 2.0 embeddings from all 8 sessions.

These settings aim to investigate whether speech from the first session alone can predict treatment outcome and whether MFCCs or wav2vec 2.0 can capture acoustic changes across sessions that reflect therapeutic progress. As described in Section 3.3, each session is segmented into multiple 5 s intervals, from which the mean and standard deviation are computed and used to construct the input feature vectors. While training and evaluation are performed at the segment level, metrics such as recall, precision, F1-score, and RMSE are computed over all segments of the held-out patient within each LOO fold. The reported results correspond to fold-wise averages across patients (mean ± standard deviation), which prevents bias from unequal segment counts and ensures comparability across subjects.

Given that Settings 3 and 4 (features from all 8 CBT sessions) consistently produced superior performance in the regression task (Section 5), these configurations were chosen for wav2vec 2.0 and the classification experiments. This selection aimed to leverage the input representations that most effectively captured treatment-related speech dynamics.

Model hyperparameters were optimized using the GridSearchCV. The search space was defined as follows: for RF, the number of estimators was set to [50, 100], maximum depth to [None, 5], and minimum samples per leaf to [1, 2]; for XGB, the number of estimators was set to [50, 100], maximum depth to [3, 5], and learning rate to [0.1, 0.2]; for SVR, the penalty parameter C was set to [1, 10], using an RBF kernel; for LR, C was explored in [0.01, 0.1, 1, 10], with penalty set to ['l2'] and solver to ['liblinear']; and for MLP, the hyperparameters included hidden-layer-sizes set to [(100,), (50, 50), (100, 50)], activation to ['relu', 'tanh'], solver to ['adam', 'sgd'], and learning-rate to ['constant', 'adaptive'].

Hyperparameter tuning was nested within each LOO iteration as follows: for each held-out patient, hyperparameters were optimized using GridSearchCV on the training fold only, with 3-fold inner cross-validation. The best model was then evaluated on the unseen test patient. This procedure ensures the unbiased estimation of generalization performance.

In terms of computational complexity for XGB, the depth determines the number of nodes in each XGB tree, which directly impacts the training cost ($O(n_{\text{samples}} \times n_{\text{features}} \times \text{depth})$). Convergence remained stable within this search space because shallow trees and moderate learning rates in XGB prevented overfitting and enabled the boosting process to achieve residual stabilization in a small number of iterations.

For MLP, the computational cost grows with the number of hidden units and epochs, approximately $O(n_{\text{samples}} \times n_{\text{features}} \times n_{\text{hidden}} \times n_{\text{epochs}})$. Convergence behavior depended mainly on the learning rate schedule (constant or adaptive) and the optimizer; Adam converged faster, while SGD required more iterations but offered more stable generalization. To mitigate divergence and local minima, cross-validation ensured the best parameter selection [28].

Experiments were run in Python 3.10 with scikit-learn v1.2, XGBoost v1.7, and PyTorch v2.0 (for wav2vec2.0 feature extraction). The wav2vec2.0 variant used was *facebook/wav2vec2-base-960h*. Class imbalance in classification tasks was handled using class_weight = balanced (for RF, LR) and scale_pos_weight (for XGB). No early stopping was applied to classical ML models, while MLP used default scikit-learn stopping with a patience of 200 epochs.

5. Results

The ML models were applied across the defined input settings (1–4) to predict depression treatment outcomes per group. Results are reported as the mean and standard

deviation of RMSE and MAE for regression tasks. For classification, we present fold-wise macro-averaged recall (M), which averages recall across classes within each fold to account for imbalance, and report the mean and standard deviation across folds for both binary and multiclass tasks. In the binary setting, we additionally report positive-class recall (P), i.e., the sensitivity for the clinically relevant class, together with precision and F1-score. Confusion matrices are row-normalized and aggregated across all folds (pooled predictions), whereas the summary metrics in the tables correspond to the fold-wise mean.

5.1. Young Adults and Adolescents

This cluster includes all 35 patients in the study. Table 2 presents the results of the regression task (R) for this group.

Table 2. Young adults and adolescents: regression task.

Experiment	Input Setting	Model	RMSE	MAE
Baseline 1	1	Dummy Regressor	0.082 ± 0.079	0.082
Baseline 2	1	Linear Regression	0.072 ± 0.066	0.072
R1	1	RF	0.071 ± 0.073	0.077
		XGB	0.074 ± 0.071	0.074
		SVR	0.084 ± 0.064	0.080
		MLP	$\mathbf{0.064 \pm 0.064}$	0.067
R2	2	RF	0.071 ± 0.073	0.077
		XGB	0.075 ± 0.073	0.074
		SVR	0.086 ± 0.069	0.082
		MLP	0.079 ± 0.077	0.080
R3	3	RF	0.071 ± 0.073	0.077
		XGB	0.073 ± 0.072	0.074
		SVR	0.085 ± 0.068	0.082
		MLP	$\mathbf{0.063 \pm 0.065}$	0.074
R4	4	RF	0.071 ± 0.073	0.076
		XGB	0.074 ± 0.072	0.074
		SVR	0.079 ± 0.069	0.079
		MLP	$\mathbf{0.057 \pm 0.049}$	0.057

The results show that MLP predicted depression with an RMSE of 0.064 using only MFCCs from the first session (R1), suggesting that early-session speech may already provide valuable information for outcome prediction. As baselines, the Dummy Regressor yielded an RMSE of 0.082, while a simple linear clinical model using age and gender achieved 0.072. In contrast, the MLP trained on the entire CBT treatment achieved an RMSE of 0.063 using MFCCs (R3) and an even lower RMSE of 0.057 with wav2vec 2.0 embeddings (R4).

Relative to the Dummy baseline, the wav2vec 2.0 model reduced RMSE by 30.5% ($\Delta = 0.025$, 95% CI $[0.007, 0.046]$, Wilcoxon $p = 0.0065$). Compared to the linear clinical baseline, the reduction was 20.8% ($\Delta = 0.015$, 95% CI $[-0.001, 0.031]$, Wilcoxon $p = 0.054$). These results indicate that the models capture informative speech features beyond simple central tendency measures and that the improvements over baselines are statistically robust and show a positive trend relative to a simple clinical baseline.

This performance is further illustrated in Figure 3, which shows the predicted versus true standardized depression scores using MLP with wav2vec 2.0 (R4). While some underestimation is observed at higher true scores, predictions generally follow the expected trend, indicating the model's ability to capture individual variation in depressive severity.

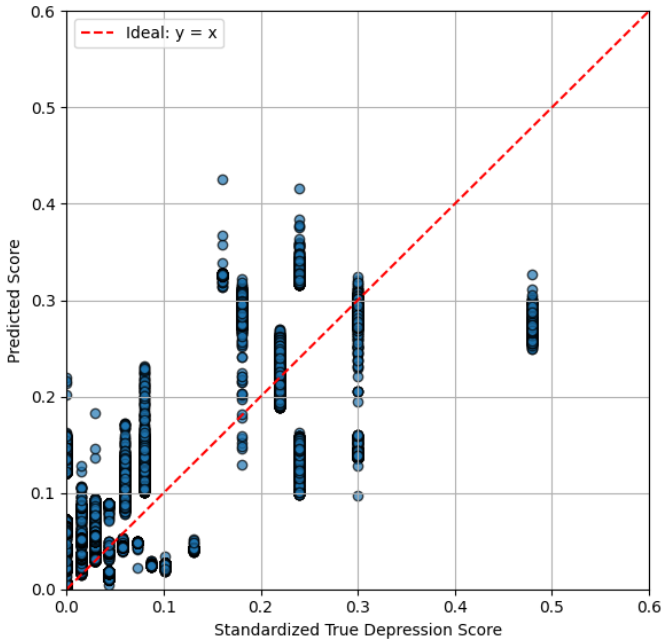

Figure 3. Predicted vs. true depression scores for wav2vec 2.0 using MLP.

To explore which features MLP relied on, we applied SHapley Additive exPlanations (SHAP) to its best-performing setup using wav2vec 2.0 embeddings (Figure 4). Certain embedding dimensions, such as Feature 2, corresponding to a specific index in the wav2vec 2.0 embedding vector, contributed more to the predictions. While these features are not semantically interpretable, the model appears to have captured stable latent speech patterns associated with treatment response. Nevertheless, to provide more clinically interpretable insights, we derived proxies from MFCC features as follows: MFCC1 as vocal intensity proxy and averages of MFCC2–5, MFCC6–9, and MFCC10–13 as spectral slope measures (reflecting spectral balance and timbre). Figure 5 illustrates group-level changes from session 1 to session 8 for Worse versus Better patients. While effect sizes were modest, 95% confidence intervals indicated systematic differences in intensity and spectral slopes across therapy, with Better patients (label = 1) showing a trend toward increased intensity and more stable spectral slopes compared to Worse patients (label = 0).

For both binary (B) and multiclass (M) tasks (Table 3), XGB yielded the best results, with a fold-wise macro-averaged recall (M) of 54.8% when using wav2vec 2.0 for the binary task and 45.7% for the multiclass task using both MFCCs and wav2vec 2.0 (Table 4). Reported summary values correspond to the mean and standard deviation across folds, whereas confusion matrices are row-normalized by true class and aggregated across all folds (pooled predictions), as illustrated in (Figure 6).

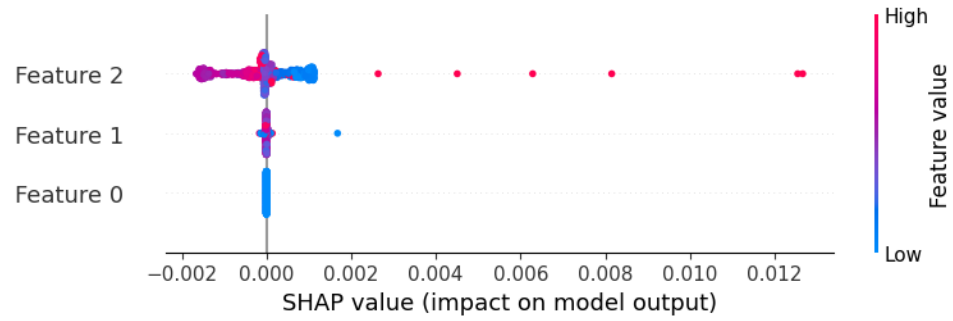

Figure 4. SHAP summary of wav2vec 2.0 feature importance using MLP.

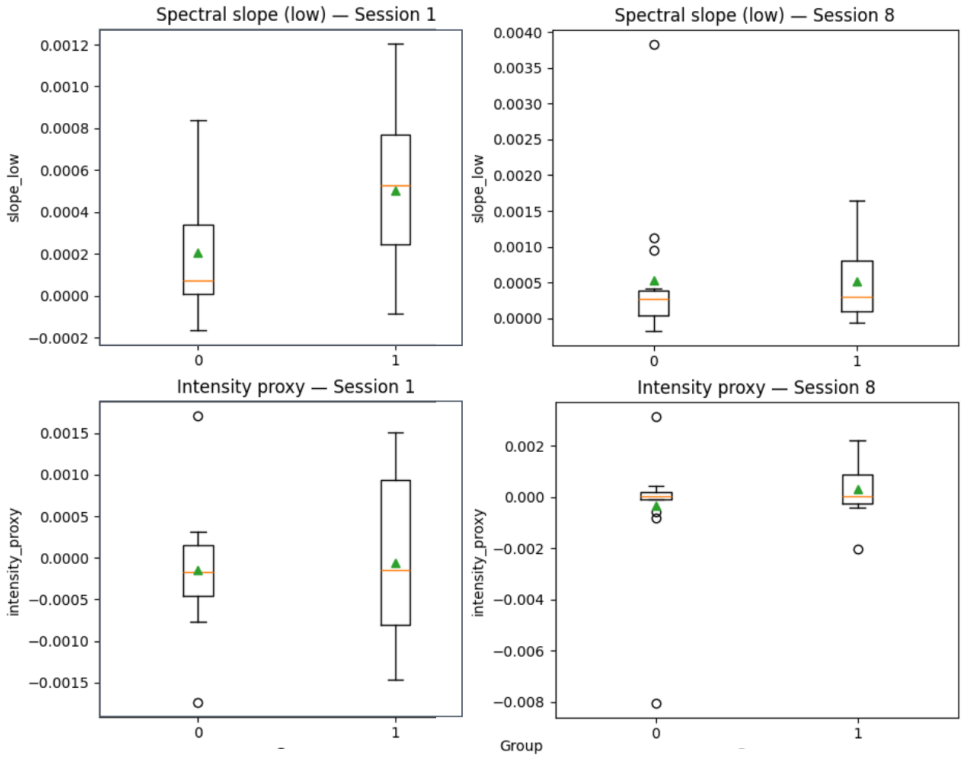

Figure 5. Changes in MFCC-derived intensity and spectral slopes from Session 1 to Session 8 in Worse (0) and Better (1) patient groups. Boxplots show medians (orange lines), means (green triangles), quartiles, and outliers (circles).

Table 3. Young adults and adolescents: binary task.

Experiment	Input	Model	Recall (M)	Recall (P)	Precision (P)	F1 (P)
B1	3	RF	0.383 ± 0.409	0.136	0.323	0.157
		LR	0.259 ± 0.194	0.232	0.452	0.277
		XGB	0.498 ± 0.483	0.194	0.194	0.194
		MLP	0.297 ± 0.261	0.191	0.420	0.235
B2	4	RF	0.371 ± 0.446	0.235	0.258	0.226
		LR	0.493 ± 0.501	0.238	0.258	0.244
		XGB	**0.548 ± 0.506**	0.194	0.194	0.194
		MLP	0.401 ± 0.468	0.164	0.322	0.167

Table 4. Young adults and adolescents: multiclass classification task.

Experiment	Input Setting	Model	Recall (M)	Precision (M)	F1 (M)
M1	3	RF	0.199 ± 0.248	0.255	0.461
		LR	0.282 ± 0.302	0.568	0.571
		XGB	**0.457 ± 0.505**	0.269	0.269
		MLP	0.303 ± 0.303	0.519	0.692
M2	4	RF	0.343 ± 0.449	0.516	0.534
		LR	0.413 ± 0.404	0.421	0.394
		XGB	**0.457 ± 0.505**	0.502	0.520
		MLP	0.388 ± 0.462	0.593	0.609

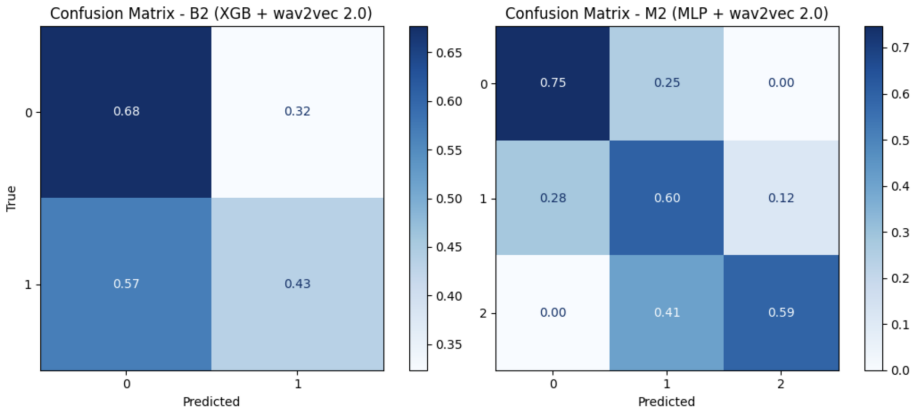

Figure 6. Young adults and adolescents: confusion matrices of the best setting.

5.2. Young Adults

When evaluating the young adults separately in the regression task (Table 5), MLP outperformed all models, achieving its best result with wav2vec 2.0 (RMSE of 0.027).

Table 5. Young adults: regression task.

Experiment	Input Setting	Model	RMSE	MAE
R1	3	RF	0.040 ± 0.037	0.042
		XGB	0.038 ± 0.035	0.038
		SVR	0.042 ± 0.020	0.042
		MLP	0.032 ± 0.022	0.035
R2	4	RF	0.040 ± 0.037	0.041
		XGB	0.037 ± 0.036	0.037
		SVR	0.042 ± 0.020	0.042
		MLP	0.031 ± 0.023	0.032
R3	3	RF	0.041 ± 0.036	0.039
		XGB	0.037 ± 0.035	0.037
		SVR	0.042 ± 0.020	0.042
		MLP	0.030 ± 0.023	0.031
R4	4	RF	0.035 ± 0.036	0.036
		XGB	0.037 ± 0.035	0.037
		SVR	0.042 ± 0.020	0.042
		MLP	**0.027 ± 0.023**	0.029

Improved performance was also observed in the binary task (Table 6), where XGB reached a macro recall of 0.7 using wav2vec 2.0 (B2). In contrast, in the multiclass task, both MFCCs and wav2vec 2.0 yielded a macro recall of 0.54 when using XGB (Table 7). Results are also illustrated in the confusion matrices (Figure 7).

Table 6. Young adults: binary task.

Experiment	Input	Model	Recall (M)	Recall (P)	Precision (P)	F1 (P)
B1	3	RF	0.368 ± 0.414	0.129	0.300	0.143
		LR	0.285 ± 0.207	0.184	0.350	0.223
		XGB	0.657 ± 0.480	0.150	0.150	0.150
		MLP	0.448 ± 0.434	0.112	0.250	0.126
B2	4	RF	0.600 ± 0.475	0.100	0.100	0.100
		LR	0.550 ± 0.510	0.168	0.200	0.177
		XGB	**0.697 ± 0.441**	0.150	0.150	0.150
		MLP	0.465 ± 0.469	0.123	0.200	0.132

Table 7. Young adults: multiclass classification task.

Experiment	Input	Model	Recall (M)	Precision (M)	F1 (M)
M1	3	RF	0.277 ± 0.354	0.562	0.492
		LR	0.171 ± 0.156	0.112	0.112
		XGB	**0.542 ± 0.509**	0.381	0.381
		MLP	0.372 ± 0.393	0.600	0.488
M2	4	RF	0.542 ± 0.509	0.449	0.444
		LR	0.388 ± 0.416	0.578	0.417
		XGB	**0.542 ± 0.509**	0.414	0.416
		MLP	0.444 ± 0.465	0.426	0.440

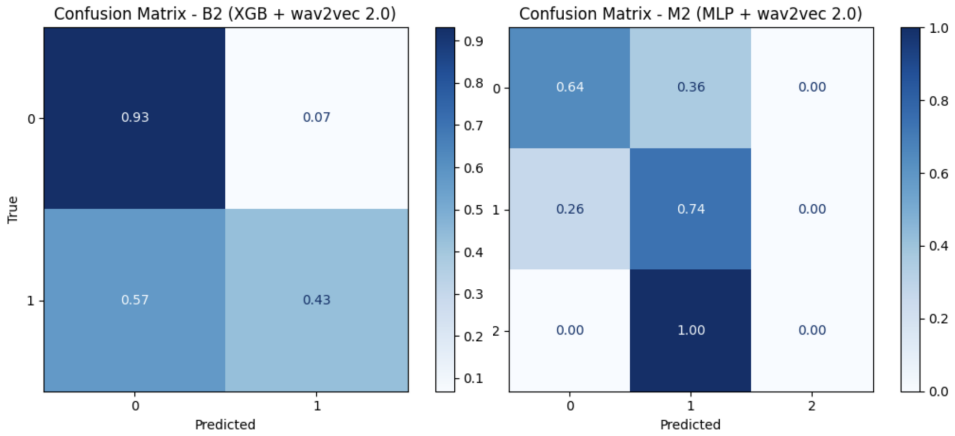

Figure 7. Young adults: confusion matrices of the best setting.

5.3. Adolescents

For the adolescents' group, the results are not as promising as those observed in the other groups, possibly due to the small sample size (11 patients). Across all tasks, MFCCs and wav2vec 2.0 yielded similar results, with very slightly improved performance when using XGB (Tables 8–10), and as illustrated in the confusion matrices (Figure 8).

Table 8. Adolescents: regression task.

Experiment	Input Setting	Model	RMSE	MAE
R1	3	RF	0.140 ± 0.085	0.140
		XGB	**0.139 ± 0.084**	0.140
		SVR	0.142 ± 0.086	0.135
		MLP	0.152 ± 0.102	0.145
R2	4	**RF**	0.140 ± 0.085	0.140
		XGB	0.140 ± 0.083	0.140
		SVR	0.149 ± 0.094	0.145
		MLP	0.154 ± 0.086	0.145
R3	3	**RF**	0.140 ± 0.085	0.140
		XGB	0.140 ± 0.084	0.140
		SVR	0.156 ± 0.083	0.145
		MLP	0.144 ± 0.095	0.133
R4	4	RF	0.140 ± 0.085	0.140
		XGB	**0.139 ± 0.084**	0.143
		SVR	0.195 ± 0.161	0.136
		MLP	0.150 ± 0.111	0.114

Table 9. Adolescents: binary task.

Experiment	Input	Model	Recall (M)	Recall (P)	Precision (P)	F1 (P)
B1	3	RF	0.189 ± 0.461	0.137	0.455	0.170
		LR	0.164 ± 0.196	0.250	0.545	0.338
		XGB	**0.364 ± 0.505**	0.273	0.273	0.273
		MLP	0.224 ± 0.462	0.253	0.545	0.323
B2	4	RF	0.316 ± 0.318	0.183	0.363	0.184
		LR	0.118 ± 0.127	0.223	0.363	0.244
		XGB	**0.364 ± 0.505**	0.272	0.272	0.272
		MLP	0.282 ± 0.292	0.195	0.206	0.363

Table 10. Adolescents: multiclass classification task.

Experiment	Input	Model	Recall (M)	Precision (M)	F1 (M)
M1	3	RF	0.080 ± 0.167	0.168	0.318
		LR	0.110 ± 0.168	0.238	0.409
		XGB	**0.273 ± 0.467**	0.387	0.392
		MLP	0.207 ± 0.307	0.172	0.363
M2	4	RF	0.134 ± 0.321	0.334	0.231
		LR	0.185 ± 0.328	0.334	0.334
		XGB	**0.237 ± 0.467**	0.387	0.392
		MLP	0.244 ± 0.402	0.334	0.231

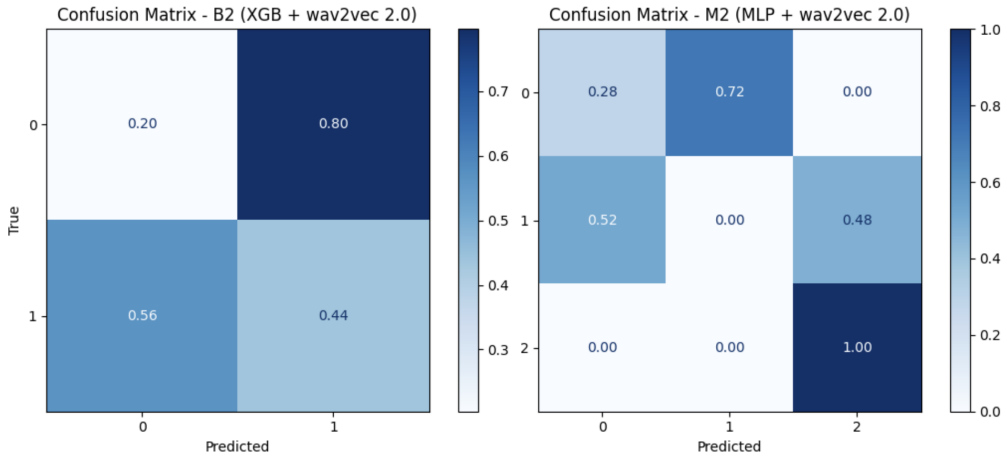

Figure 8. Adolescents: confusion matrices of the best setting.

6. Discussion

The proposed models aim to predict changes in depression following CBT, a clinically relevant treatment outcome. Using speech from all CBT sessions, the MLP with wav2vec 2.0 achieved an RMSE of 0.057, outperforming the baseline. In the binary task, it reached a recall of 70% for detecting improvement versus worsening in young adults. These results suggest that speech-based features, particularly deep acoustic embeddings, can provide valuable insight into treatment response, potentially supporting clinicians in monitoring progress and adapting interventions in real time.

Although CBT is effective for many patients, some do not respond as expected. Our dataset included participants whose depression scores increased after treatment, which is consistent with real-world clinical variability. These cases highlight the importance of predictive tools for identifying early signs of non-response. These tools allow clinicians

to modify the therapeutic plan by intensifying treatment, introducing complementary strategies, or reevaluating the care approach to further improve the treatment process.

A key strength of this work is its clinical integration. Unlike most studies that use static speech data, our approach uses real-world, clinician-guided CBT sessions to model changes in depression over time. Given this, direct comparison with existing studies is challenging due to differences in clinical context, target outcomes, and data collection protocols. Nevertheless, our method addresses the following understudied need: predicting treatment response dynamically from longitudinal speech data.

Limitations

Despite these contributions, limitations remain—particularly for the adolescent group, which relied on the CDI, a self-reported questionnaire that may introduce subjectivity in depression scoring. Combined with the limited sample size, it may have led to a higher concentration of severe cases and impacted the generalizability of the models in this group.

We acknowledge that this is an exploratory study based on a limited cohort, which is in line with dataset sizes commonly reported in speech-based depression recognition research. As such, the results are not sufficient for clinical deployment but provide an important step toward validating speech as a biomarker of depression in a real-world clinical context. As the clinical trial is still ongoing, future work will focus on replicating the approach with a larger sample size and conducting external validation to establish stability and generalizability. We also intend to incorporate recurrent models, such as LSTMs or Transformers, capable of capturing temporal dependencies in longitudinal speech data, which may enhance the modeling of symptom evolution across therapy sessions.

Future work will further strengthen robustness and clinical interpretability. Specifically, we intend to carry out the following: (i) integrate calibration methods (e.g., calibration curves, Brier score, and decision-curve analysis); (ii) perform longitudinal SHAP analyses to assess whether the same embedding dimensions consistently drive predictions across therapy; (iii) include preprocessing enhancements (loudness normalization and silence handling) and robustness checks with alternative segment durations; (iv) evaluate diarization robustness by reporting objective metrics (DER/JER) and testing VAD with role-based aggregation to validate patient–therapist separation; and (v) explore patient-level aggregation strategies (e.g., majority vote and probability averaging) to complement segment-level evaluation and further improve clinical interpretability.

7. Conclusions

This study demonstrates the potential of speech-based features, particularly wav2vec 2.0, combined with ML models for predicting MDD in hikikomori patients undergoing a CBT intervention. The results indicate that speech captured during the initial session can provide clinically relevant information for estimating depression treatment outcomes, whereas incorporating all sessions' data enhances the modeling of depression progression. This suggests that speech may serve as a valuable digital biomarker to support early clinical decisions and timely interventions.

Author Contributions: Conceptualization, S.S.L.; methodology, S.S.L., R.S., and S.N.; software, S.S.L., R.S., and S.N.; validation, M.G.R., M.B., and A.T.; formal analysis, S.S.L. and S.N.; investigation, S.S.L. and S.N.; resources, R.S. and S.N.; data curation, M.G.R., M.B., and A.T.; writing—original draft preparation, S.S.L.; writing—review and editing, S.S.L., S.N., and R.S.; visualization, S.S.L.; supervision, R.S.; project administration, R.S.; funding acquisition, R.S. and M.B. All authors have read and agreed to the published version of the manuscript.

Funding: Part of this research was supported by the project SOLITAIRE—"Digital interventions for Social isOLation In youThs And theIR familiEs", funded by the European Union—Next Generation EU—NRRP M6C2—Investment 2.1 "Enhancement and strengthening of biomedical research in the NHS". Project number: PNRR-MAD-2022-12376834, CUP: G43C22004010006.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration 468 of Helsinki and approved by the Ethical Committees "AREA SUD OVEST VENETO", operating from the University of Verona, and of the University of Milan. Clinical interviews and telematic CBT sessions were carried out through the COD20 telemedicine platform developed by the University of Milan, designed for patient–specialist video consultations. The platform complies with the European privacy regulation policy (Article 13 of EU Regulation 679/2016 for the Protection of Personal Data and Article 15 et seq., EU Regulation 2016/679) and follows AWS Security standards to minimize the risk of inappropriate data access.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The authors are willing to share the extracted features upon reasonable request.

Acknowledgments: We would like to thank the SOLITAIRE group for their contribution (Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy and Azienda Ospedaliera Universitaria Integrata Verona, Verona, Italy: Mirella Ruggeri, Marcella Bellani, Maria Gloria Rossetti, Cinzia Perlini, Francesca Girelli, Niccolò Zovetti, and Maria Diletta Buio; Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy: Paolo Brambilla, Cinzia Bressi, Antonella delle Fave, and Virginia Pupi; CNR Institute of Neuroscience, Vedano al Lambro, Italy: Fabrizia Guarnieri and Edoardo Moretto; Department of Computer Science, University of Milan, Milan, Italy: Roberto Sassi, Maria Renata Guarneri, Stavros Ntalampiras, and Samara Soares Leal; Unit for Severe Disabilities in Developmental Age and Young Adults, Associazione La Nostra Famiglia-IRCCS E. Medea, Scientific Hospital for Neurorehabilitation, Brindisi, Italy: Isabella Fanizza, Lara Scialpi, Giorgia Carlucci, and Mariangela Leucci; Scientific Institute IRCCS Eugenio Medea, Scientific Direction, Bosisio Parini, Lecco, Italy: Antonio Trabacca).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
SDR	Speech Depression Recognition
RF	Random Forest
SVR	Support Vector Regression
LR	Logistic Regression
XGB	Extreme Gradient Boosting
MLP	Multi-Layer Perceptron
MFCCs	Mel-Frequency Cepstral Coefficients
MDD	Major Depression Disorder
CBT	Cognitive Behavioral Therapy
HDRS	Hamilton Depression Rating Scale
CDI	Children's Depression Inventory

References

1. Malhi, V. Depression. *Lancet* **2018**, *392*, 2299–2312. [CrossRef]
2. Vasha, Z.N.; Sharma, B.; Esha, I.J.; Al Nahian, J.; Polin, J.A. Depression detection in social media comments data using machine learning algorithms. *Bull. Electr. Eng. Inform.* **2023**, *12*, 987–996. [CrossRef]
3. OECD/European Observatory on Health Systems and Policies. *Italy: Country Health Profile, State of Health in the EU*; OECD Publishing: Brussels, Belgium, 2023; ISBN 978926477151.

4. Oscoz-Irurozqui, M.; Villani, L.; Martinelli, S.; Ricciardi, W.; Gualano, M.R. Trend analysis of antidepressant consumption in Italy from 2008 to 2022 in a public health perspective. *Sci. Rep.* **2025**, *15*, 12124. [CrossRef]
5. Kato, T.A.; Kanba, S.; Teo, A.R. Hikikomori: Multidimensional understanding, assessment, and future international perspectives. *Psychiatry Clin. Neurosci.* **2019**, *73*, 427–440. [CrossRef] [PubMed]
6. Fava, G.A.; Kellner, R.; Munari, F.; Pavan, L. The Hamilton depression rating scale in normals and depressives. *Acta Psychiatr. Scand.* **1982**, *66*, 26–32. [CrossRef]
7. Helsel, W.J.; Matson, J.L. The assessment of depression in children: The internal structure of the child depression inventory (CDI). *Behav. Res. Ther.* **1984**, *22*, 289–298. [CrossRef] [PubMed]
8. Iyortsuun, N.K.; Kim, S.-H.; Yang, H.-J.; Kim, S.-W.; Jhon, M. Additive cross-modal attention network (ACMA) for depression detection based on audio and textual features. *Electronics* **2024**, *12*, 20479–20489. [CrossRef]
9. Hillhouse, T.M.; Porter, J.H. A brief history of the development of antidepressant drugs: From monoamines to glutamate. *Exp. Clin. Psychopharmacol.* **2015**, *23*, 1. [CrossRef]
10. Wu, P.; Wang, R.; Lin, H.; Zhang, F.; Tu, J.; Sun, M. Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Trans. Intell. Technol.* **2023**, *8*, 701–711. [CrossRef]
11. Ngiam, K.Y.; Khor, W. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* **2019**, *20*, e262–e273. [CrossRef]
12. Alghifari, M.F.; Gunawan, T.S.; Kartiwi, M. Development of sorrow analysis dataset for speech depression prediction. In Proceedings of the 2023 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Kuala Lumpur, Malaysia, 22–25 May 2023; pp. 1–6.
13. Aloshban, N.; Esposito, A.; Vinciarelli, A. What you say or how you say it? Depression detection through joint modeling of linguistic and acoustic aspects of speech. *Cogn. Comput.* **2022**, *14*, 1585–1598. [CrossRef]
14. Low, D.M.; Bentley, K.H.; Ghosh, S.S. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investig. Otolaryngol.* **2020**, *5*, 96–116. [CrossRef]
15. Espinola, C.W.; Gomes, J.C.; Pereira, J.M.S.; Dos Santos, W.P. Detection of major depressive disorder using vocal acoustic analysis and machine learning—An exploratory study. *Res. Biomed. Eng.* **2021**, *37*, 53–64. [CrossRef]
16. Leal, S.S.; Ntalampiras, S.; Sassi, R. Speech-based depression assessment: A comprehensive survey. *IEEE Trans. Affect. Comput.* **2024**, *16*, 1318–1333. [CrossRef]
17. Martinengo, L.; Van Galen, L.; Lum, E.; Kowalski, M.; Subramaniam, M.; Car, J. Suicide prevention and depression apps' suicide risk assessment and management: A systematic assessment of adherence to clinical guidelines. *BMC Med.* **2019**, *17*, 1–12. [CrossRef]
18. Rossetti, M.G.; Perlini, C.; Girelli, F.; Zovetti, N.; Brambilla, P.; Bressi, C.; Bellani, M. Developing a brief telematic cognitive behavioral therapy for the treatment of social isolation in young adults. *Front. Psychol.* **2024**, *15*, 1433108. [CrossRef]
19. Verde, L.; Raimo, G.; Vitale, F.; Carbonaro, B.; Cordasco, G.; Marrone, S.; Esposito, A. A lightweight machine learning approach to detect depression from speech analysis. In Proceedings of the IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), Baltimore, MD, USA, 1–3 November 2021; pp. 330–335.
20. Zhang, X.; Zhang, X.; Chen, W.; Li, C.; Yu, C. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. *Sci. Rep.* **2024**, *14*, 1. [CrossRef] [PubMed]
21. Cummins, N.; Scherer, S.; Krajewski, J.; Schnieder, S.; Epps, J.; Quatieri, T.F. A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* **2015**, *71*, 10–49. [CrossRef]
22. Huang, Z.; Epps, J.; Joachim, D.; Chen, M. Depression detection from short utterances via diverse smartphones in natural environmental conditions. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 3393–3397.
23. Huang, X.; Wang, F.; Gao, Y.; Liao, Y.; Zhang, W.; Zhang, L.; Xu, Z. Depression recognition using voice-based pre-training model. *Sci. Rep.* **2024**, *14*, 1. [CrossRef]
24. Ambady, N.; Bernieri, F.J.; Richeson, J.A. Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Adv. Exp. Soc. Psychol.* **2000**, *32*, 201–271.
25. Yin, F.; Du, J.; Xu, X.; Zhao, L. Depression detection in speech using transformer and parallel convolutional neural networks. *Electronics* **2023**, *12*, 328. [CrossRef]
26. Ntalampiras, S. Model ensemble for predicting heart and respiration rate from speech. *IEEE Intell. Comput.* **2023**, *27*, 15–20. [CrossRef]
27. Ntalampiras, S. Toward language-agnostic speech emotion recognition. *J. Audio Eng. Soc.* **2020**, *68*, 7–13. [CrossRef]
28. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv* **2020**, arXiv:2006.11477.

29. Cawley, G.C.; Talbot, N.L.C. Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers. *Pattern Recognit.* **2003**, *36*, 2585–2592. [CrossRef]
30. Mohammadi, M.; Al-Azab, F.; Raahemi, B.; Richards, G.; Jaworska, N.; Smith, D.; Salle, S.d.; Blier, P.; Knott, V. Data mining EEG signals in depression for their diagnostic value. *BMC Med. Inform. Decis. Mak.* **2015**, *15*, 1. [CrossRef]