Brazilian Journal of Psychiatry
bjp
Revista Brasileira de Psiquiatria

Braz J Psychiatry. 2026;48:e20254419
doi:10.47626/1516-4446-2025-4419
Brazilian Psychiatric Association

ORIGINAL ARTICLE

AI-detected auditory findings of depression and suicide risk

Sena **Ozden**,[1] Mustafa **Sert**,[2,3] Şakir **Gica**,[4] Orhan **Cinar**,[5] Irmak Uzun **Bayar**,[6] Ayhan Yahya **Acar**,[7] Mehmet **Ak**[8]

[1]*Department of Psychiatry, Kızıltepe State Hospital, Mardin, Turkey.* [2]*Department of Electrical Engineering and Computer Science, University of California, Irvine, CA, USA.* [3]*Department of Computer Engineering, Baskent University, Ankara, Turkey.* [4]*Department of Psychiatry, Necmettin Erbakan University, Konya, Turkey.* [5]*Department of Emergency Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey.* [6]*Department of Management Information Systems, Ankara Bilim University, Ankara, Turkey.* [7]*Department of Pathology and Clinical Sciences Education, Mercer University, School of Medicine, Macon, GA, USA.* [8]*Department of Psychiatry, Necmettin Erbakan University, Konya, Turkey.*

Objective: Suicide is a major global health concern and one of the leading causes of preventable death. Currently, there is a lack of objective data to assess suicide risk in individuals with depression. This study explores the potential of voice analysis as an objective tool for suicide risk assessment and improved diagnostic accuracy.
Methods: Ninety participants divided into three groups (Near-Term Suicidal, Depressed, and Control) provided audio recordings of standardized text readings. Mel-frequency cepstral coefficients, deep learning features (VGGish), formants, and prosodic features were extracted and analyzed using a machine learning model.
Results: Among the analyzed voice features, Mel-frequency cepstral coefficients were more successful for the "high suicide risk or not" and "depression or high suicide risk" tasks, with an accuracy of 90.0% and 68.3%, respectively. For the "depression or not" task, VGGish representation achieved an accuracy of 73.3%.
Conclusions: To our knowledge, this is the first study to employ VGGish features in the suicidality assessment. The findings demonstrate significant differences in vocal parameters across varying suicide risk levels, supporting the potential of voice analysis as a biomarker for suicide risk in depression. Given its non-invasive nature and real-time integration capability, voice analysis offers a promising approach for clinical applications.

Keywords: Suicide; voice analysis; artificial intelligence; classification; machine learning

Introduction

Suicide, a complex public health problem that claims over 700,000 lives each year (which equates to one person every 40 seconds), is the fourth leading cause of death among individuals aged 15-29 years.[1] It remains a leading cause of preventable death in both developing and developed countries, affecting people of all genders, ages, and ethnic backgrounds. Approximately 90% of those who attempt suicide have a diagnosable psychiatric disorder, with depression, bipolar disorder, schizophrenia, and substance use disorders being the most common.[2,3] Of these, depression is the most strongly associated with an increased risk of suicide.[2] Individuals with depression are 20 times more likely to die by suicide than those without mood disorders.[4] The heterogeneity of suicidal behaviors in clinical presentation and treatment highlights the need for objective biomarkers that can enhance risk prediction and complement traditional, subjective methods still prevalent in psychiatric practice.[5-7] Recent studies have linked suicidal behavior to abnormalities in the serotonergic system, inflammation, lipids, and endocannabinoids, yet no predictive biomarkers have been identified.[8]

Although 77% of individuals who die by suicide have had contact with health care services within the preceding year and 45% within the final month,[9,10] risk assessment still relies heavily on self-reporting. As suicidal ideation is often denied even by those who attempt or complete suicide,[11] this subjectivity precludes accurate risk detection. Thus, objective and scalable assessment methods that transcend patient self-disclosure are needed.

Speech serves as a fundamental data source for diagnosing and treating mental disorders, providing clinicians with critical insights into cognitive processes and emotional states.[12] Psychological states are known to

Correspondence: Sena Ozden, Department of Psychiatry, Kızıltepe State Hospital, 47400 Kızıltepe, Mardin, Turkey.
E-mail: yunden_sena@hotmail.com
Submitted Jul 08 2025, accepted Sep 29 2025.

How to cite this article: Ozden S, Sert M, Gica S, Cinar O, Bayar IU, Acar AY, et al. AI-detected auditory findings of depression and suicide risk. Braz J Psychiatry. 2026;48:e20254419. http://doi.org/10.47626/1516-4446-2025-4419

influence speech production, a phenomenon long recognized by clinicians.[13] Kraepelin first described altered vocal patterns in depression, such as reductions in pitch (auditory perception of tone), intensity, tempo, and vocal variation.[14] Recent studies have also identified speech latency as an objective marker of psychomotor slowing and as a predictor of both diagnostic status and treatment response.[15]

Emerging evidence highlights voice analysis as a promising biomarker for suicide risk.[13,16] Silverman reported tonal changes in suicidal speech that are distinct from depression.[17] Acoustic analyses of suicidal speech have been conducted,[18,19] and recent studies have applied machine learning for classification.[20,21] France et al.[18] reported 80% accuracy in distinguishing suicidal from control male voices using formant and spectral features. Similarly, F2 (second formant) characteristics and spectral features could effectively distinguish depressive from pre-suicidal speech, with an accuracy of up to 90% for reading and 85.75% for interviews.[22] Acoustic features from audio recordings of U.S. veterans have been used to classify suicidal utterances, achieving an accuracy rate of 73%.[20]

Voice analysis is a strong candidate for suicide risk prediction and is cost-effective, non-invasive, and practical in clinical settings.[13] Voice data can be easily collected by microphones, mobile phones, or wearable devices, allowing data collection during face-to-face sessions and remotely via telepsychiatry, which facilitates real-time, personalized analysis using machine learning tools. Voice analysis may reduce barriers such as cost, stigma, and limited access by supporting continuous symptom monitoring and facilitating early identification of at-risk individuals.[23]

The current study investigated the potential of voice analysis as an objective tool for assessing suicide risk in depressed individuals. Classification models were developed using acoustic and deep features, including mel-frequency cepstral coefficients (MFCCs), VGGish, formants, and prosody. To our knowledge, this is the first study to employ VGGish features in suicidality assessment. We hypothesized that machine learning models trained on vocal features would yield high classification performance and support clinical decision-making in suicide risk evaluation.

Methods

Study participants and setting

Sixty patients with major depressive disorder and 30 healthy controls aged 18-65 years who presented to the clinic at (blinded for reviewer) between January 2022 and January 2023 were recruited. The study sample was divided into three groups: a near-term suicidal group (NTSG), consisting of 30 patients who met the DSM-5 criteria for major depressive disorder and had attempted suicide within the preceding 10 days; a depressed group (DG), also consisting of 30 patients who met the DSM-5 criteria for major depressive disorder but had neither active suicidal thoughts nor a history of suicide attempts within the past year; and a control group (CG) consisting of 30 individuals with no known psychiatric diagnosis or history of medication use. All participants provided written informed consent. Exclusion criteria were mental impairment, medical conditions that could affect the interview, $<$ 5 years of formal education, alcohol or substance use disorder, electroconvulsive therapy within the past year, known hearing impairments, psychotic disorders, or autism spectrum disorders. In addition, participants experiencing depressive episodes with psychotic features or mixed affective states were also excluded to ensure diagnostic homogeneity.

The sample size was calculated at $\geqslant$ 90 using G Power 3.1.9.2, assuming an effect size of 0.36 and a power of 0.85.

Data collection tools

The Hamilton Depression Rating Scale (HDRS) is a clinician-administered instrument used to assess depression severity. The 24-item version provides a more comprehensive evaluation than the original 17-item scale.[24,25] Its Turkish validity and reliability were established by Akdemir et al.[26]

The Hamilton Anxiety Rating Scale (HARS) is a semi-structured tool commonly used to assess anxiety severity.[27] The Turkish version was validated by Yazıcı et al.[28]

The Beck Hopelessness Scale (BHS) is a 20-item self-report measure to evaluate negative expectations about the future.[29] Its Turkish validity and reliability were established by Seber et al.[30]

The Young Mania Rating Scale (YMRS), assesses manic symptom severity based on clinical observation and patient report.[31] Its Turkish validity and reliability were confirmed by Karadağ et al.[32]

Clinical assessments

All patients had previously received a clinical diagnosis of major depressive disorder. This diagnosis was reconfirmed by a single experienced clinician, who also administered the standardized assessments (HDRS, HARS, and YMRS) to ensure consistency. The YMRS was used to exclude manic or mixed episodes, and hopelessness was assessed via self-report using the BHS.

Preprocessing and voice feature extraction

As no validated Turkish reading material currently exists, the participants read aloud a 2-3-minute text developed for this study. Recordings were made in a quiet room using an Apple iPhone 12 with a triple microphone array, with participants maintaining a consistent distance from the device. The text was designed to include varied phonetic and expressive elements to facilitate robust extraction of relevant vocal features.

Preprocessing

The audio data were converted to a 16-bit PCM mono WAV format. All speech recordings were digitized at a sampling rate of 44,100 Hz with 16-bit resolution. Silent portions at the beginning and end of the recordings were removed. The recordings were stored anonymously.

Prosodic features

Prosodic features capture rhythm, stress, and intonation variations in speech, including pitch, loudness, speaking rate, energy dynamics, and F0.[13] We focused on pitch and loudness, which are robust indicators of emotional states and reflect psychological experiences related to depression and suicidality.[33] Depressed individuals often show a lower pitch and reduced variability, indicating hopelessness and emotional blunting. Loudness reduction, linked to psychomotor retardation or withdrawal, also correlates with suicidal ideation.[13,34]

Formant features

Formants characterize the vocal tract's resonant properties and contribute to the spectral profile of speech. We extracted F1-F3 frequencies due to their relevance to articulation and phonetic quality.

Mel-frequency cepstral coefficients

MFCCs, derived from the speech signal's frequency spectrum, simulate human auditory perception and are widely used to capture spectral-temporal patterns. In this study, 20 coefficients per time frame were extracted to represent key acoustic dynamics.[35,36]

Deep learning features

Deep learning features were extracted using the pre-trained VGGish model, which was trained on AudioSet, a large-scale dataset comprising over two million 10-second YouTube clips annotated with 632 sound events.[37] The model captures both acoustic and semantic properties of audio, making it suitable for complex feature representation in this study.

Machine learning classification model

Tasks defined through pairwise comparisons of the three study groups were created to evaluate the spectrum of suicide risk, including a healthy mental state, depression, and high suicide risk, using an artificial intelligence model at each stage. The CG and the DG were compared for the "depression or not" task, whereas the cg and the ntsg groups were compared for the "high suicide risk or not" task. finally, the dg and the ntsg were compared for the "depression or high suicide risk" task. The relevant audio features were extracted, namely "prosodic," "acoustic," and "deep learning features" (VGGish), which capture the spectral, temporal, and prosodic characteristics of the sound. Next, a machine learning model was built, which used these audio features as an input for the classification model.

For the task of distinguishing between "depression" and "suicide risk," a support vector machine with a radial basis function kernel was chosen, as it is particularly effective for high-dimensional data and non-linear decision boundaries, which is often the case with complex audio signals.[38]

The goal was to find an optimal hyperplane that could separate the two classes while maximizing the margin between them. Given a set of audio-derived feature vectors $x_i \in R^d$ (where d is the feature space dimension) and corresponding binary class labels $y_i \in \{-1, +1\}$, the support vector machine optimization problem seeks to minimize classification error while maximizing the margin. The decision function for classification was expressed as follows:

$$f(x) = sign\left(\sum\nolimits_{i=1}^{N} \alpha_i \, y_i \, K(x_i, \, x) + b\right)$$

where $K(x_i, x)$ is the radial basis function kernel function, given by:

$$(x_i, \, x) = e(-\gamma \|x_i - x\|^2)$$

This kernel function calculates the similarity between feature vectors x_i and a new test sample x, with γ controlling the width of the kernel. The optimization objective of the support vector machine is to maximize the margin between the two classes by solving the dual optimization problem, which is formulated as:

$$\alpha\left(\sum\nolimits^{N} \alpha - 1 - \sum\nolimits^{N} \alpha \, \alpha \, y \, y \, (\, , x)\right)$$

$$i = 1 \; i_2 \; j = 1 \; i \; j \; i \; j \; i$$

subject to the constraints:

$$0 \leqslant \alpha_i \leqslant C \text{ and } \sum\nolimits^{N} \alpha_i \, y_i = 0$$

$$i = 1$$

where C is the regularization parameter that controls the trade-off between margin maximization and misclassification minimization. Hyperparameter tuning was carried out by using the GridSearchCV, where values for C (ranging from 1 to 1,000) and γ (ranging from 10^{-8} to 10^{-1}) were systematically evaluated through 10-fold cross-validation. The best combination of parameters was selected based on the classification performance metrics, such as accuracy, precision, and recall.

Ten-fold cross-validation was performed on the entire dataset. In each fold, the model was trained on nine of the 10 subsets, and the remaining subset was used for testing. This process was repeated 10 times, with each fold serving as the test set once. The performance metrics presented in Table 1 are the mean values obtained from the 10-fold cross-validation. Hyperparameter tuning was performed using GridSearchCV with 10-fold cross-validation. The regularization parameter (C) was tested across a range of values from 1 to 1,000, and the kernel coefficient (γ) was evaluated in the range of 10-8 to 10-1. These hyperparameter values were systematically optimized during cross-validation to ensure the best performance for the support vector machine model.

Implementation

The implementation was written in Python and leveraged the scikit-learn library for model training, cross-validation, and hyperparameter optimization, which ensured an

Table 1 Machine learning model prediction performance

Task	Accuracy				Precision				Recall, TPR, Sensitivity			
	MFCC	VGGish	Formants	L + P	MFCC	VGGish	Formants	L + P	MFCC	VGGish	Formants	L + P
Depression or not	0.617	**0.733**	0.533	0.550	0.550	**0.810**	0.452	0.417	0.567	**0.700**	0.500	0.467
High suicide risk or not	**0.900**	0.833	0.483	0.467	**0.883**	0.825	0.485	0.333	**0.933**	0.833	0.600	0.300
Depression or high suicide risk	**0.683**	0.667	0.517	0.600	**0.662**	0.652	0.510	0.510	**0.767**	0.700	0.500	0.600

Table 1 Continued

Task	F1				TNR, Specificity			
	MFCC	VGGish	Formants	L + P	MFCC	VGGish	Formants	L + P
Depression or not	0.549	**0.716**	0.462	0.430	0.667	**0.767**	0.567	0.633
High suicide risk or not	**0.904**	0.825	0.519	0.308	**0.866**	**0.833**	0.367	0.633
Depression or high suicide risk	**0.700**	0.665	0.467	0.539	0.600	0.633	0.533	0.600

Bold type indicates the highest value among acoustic feature sets for each metric.
L + P = loudness and pitch; MFCC = mel-frequency cepstral coefficients; TNR = true negative rate; TPR = true positive rate; VGGish = deep learning features.

efficient and scalable solution. For reproducibility purposes, the analysis was performed using Python version 3.12.2, with the following libraries: scikit-learn 1.5.1, numpy 1.26.4, pandas 2.3.1, and other relevant packages.

Statistical analysis

Statistical analyses were conducted in SPSS 20. Normality was assessed via visual (histograms, probability plots) and analytical tests (Kolmogorov-Smirnov, Shapiro-Wilk). Descriptive data were reported as means (SD) for continuous variables and frequencies (%) for categorical variables. Group comparisons were performed using chi-square or Fisher's exact test for categorical variables, one-way analysis of variance for normally distributed data, and Kruskal-Wallis test for non-parametric data. Homogeneity of variance was tested to determine the appropriate post-hoc methods. A Bonferroni-corrected Mann-Whitney *U* test was used for non-parametric pairwise comparisons. $P < 0.05$ was considered statistically significant.

Ethics statement

The study was approved by the clinical research ethics committee of (blinded for reviewer) (approval number 2021/3501) and was conducted in accordance with the 1964 Declaration of Helsinki and its later amendments.

Results

Demographic and clinical characteristics of the patients

The sociodemographic and clinical characteristics of the three groups are presented in Table 2. There was a significant difference in age ($p < 0.001$), with the DG being older than both the CG ($p = 0.001$) and NTSG ($p < 0.001$). Significant differences were also found in cohabitation ($p = 0.004$), employment and financial status ($p < 0.001$), mainly between the CG and the two clinical groups. Moreover, non-psychiatric comorbidities were more prevalent in the DG and NTSG than the CG ($p = 0.003$).

HDRS, HARS, and BHS scores were compared across the three groups (Table 3). There were significant group-level differences in HDRS scores ($p < 0.001$), with mean scores of 1.30 (SD, 1.74) in the CG, 25.10 (SD, 6.26) in the DG, and 31.50 (SD, 8.71) in the NTSG. Pairwise comparisons confirmed significant differences between all three groups ($p < 0.001$ for CG vs. DG and CG vs. NTSG; $p = 0.001$ for DG vs. NTSG). The HARS scores also varied significantly between the groups ($p < 0.001$), with means of 1.17 (SD, 1.62) (CG), 14.37 (SD, 6.18) (DG), and 16.60 (SD, 7.98) (NTSG). Post-hoc analyses revealed significant differences between the CG and each clinical group ($p < 0.001$), but not between the DG and the NTSG ($p = 0.190$).

Similarly, BHS scores differed significantly across the groups ($p = 0.002$), with mean values of 9.33 (SD, 1.30), 10.90 (SD, 2.40), and 10.83 (SD, 1.88) for the CG, DG, and NTSG, respectively. Pairwise comparisons showed significant differences between the CG and both clinical groups ($p = 0.004$ and $p = 0.002$, respectively), yet no significant difference was observed between the DG and NTSG ($p = 0.976$).

In accordance with the study's inclusion criteria, participants in the CG did not use psychiatric medication, whereas those in the clinical groups were under pharmacological treatment. However, comparative analyses of psychotropic medication classes, including lithium, antidepressants, antipsychotics, benzodiazepines, and other mood stabilizers (e.g., valproic acid, carbamazepine, lamotrigine), revealed no statistically significant differences in medication use between the DG and NTSG ($p > 0.05$).

Machine learning model prediction performance

The support vector machine classifier was trained separately with MFCC, VGGish, formant and prosodic features, and the results were compared. Evaluations were performed using the 10-fold cross-validation method.

Table 2 Comparison of sociodemographic and clinical characteristics among the CG, DG, and NTSG

	CG (n=30)	DG (n=30)	NTSG (n=30)	df	p-value	p-value†	p-value‡	p-value§
Age (years)‖ (median ± SD)	30.40±7.86	42.04±13.02	28.84±9.80	2	< **0.001**	**0.001**	0.173	< **0.001**
Number of children‖ (median ± SD)	1.00±1.31	1.57±1.28	0.97±1.30	2	0.090			
Sex¶				2	0.366			
Male	14 (46.7)	10 (33.3)	9 (30.0)					
Female	16 (53.3)	20 (66.7)	21 (70.0)					
Marital status¶				2	0.076			
Single or divorced	15 (50.0)	14 (46.7)	22 (73.3)					
Married	15 (50.0)	16 (53.3)	8 (26.7)					
Cohabitants¶				4	**0.004**	0.063	**0.009**	0.179
Alone	9 (30.0)	2 (6.7)	1 (3.3)					
Family	19 (63.3)	26 (86.7)	22 (73.3)					
Others	2 (6.7)	2 (6.7)	7 (23.3)					
Education level¶				2	0.071			
Below high school	5 (16.7)	13 (43.3)	11 (36.7)					
High school or above	25 (83.3)	17 (56.7)	19 (63.3)					
Employment status¶				2	< **0.001**	< **0.001**	**0.002**	0.584
Employed	23 (76.7)	9 (30.0)‖	11 (36.7)‖¶					
Unemployed	7 (23.3)	21 (70.0)‖	19 (63.3)‖					
Financial status¶ (TL)				4	< **0.001**	**0.001**	**0.008**	0.085
Poor (< 8,500)	15 (50.0)	19 (63.3)	25 (83.3)					
Average (8,500-17,000)	3 (10.0)	10 (33.3)	3 (10.0)					
Good (> 17,000)	12 (40.0)	1 (3.3)	2 (6.7)					
Medical Comorbidity				2	0.003			
Yes	0 (0.0)	7 (23.3)	10 (33.3)					
No	30 (100.0)	23 (76.7)	20 (66.7)					

Data presented as n (%), unless otherwise specified.
For pairwise comparisons of data showing significant differences, the Bonferroni-corrected p-value was set at 0.017. For pairwise comparisons of data showing significant differences in the Kruskal-Wallis test, the Mann-Whitney *U* test was applied.
Bold type denotes statistical significance (p ⩽ 0.05).
CG = Control Group; df = degrees of freedom; DG = Depressed Group; NTSG = Near-Term Suicidal Group; TL = Turkish Lira.
† Comparison between CG and DG.
‡ Comparison between CG and NTSG.
§ Comparison between the DG and NTSG.
‖ Kruskal-Wallis test.
¶ Chi-square test.

Table 3 Comparison of HDRS, HARS, and BHS scores among the CG, DG, and NTSG

	CG (n=30)	DG (n=30)	NTSG (n=30)	Z	p-value	p-value†	p-value‡	p-value§
HDRS	1.30±1.74	25.10±6.26	31.50±8.71	64.511	< **0.001**	< **0.001**	< **0.001**	**0.001**
HARS	1.17±1.62	14.37±6.18	16.60±7.98	58.503	< **0.001**	< **0.001**	< **0.001**	0.190
BHS	9.33±1.30	10.90±2.40	10.83±1.88	12.135	**0.002**	**0.004**	**0.002**	0.976

Data presented as mean ± SD, unless otherwise specified.
Bold type denotes statistical significance (p ⩽ 0.05) according to the Kruskal-Wallis test.
For pairwise comparisons of data showing significant differences in the Kruskal-Wallis test, Bonferroni-corrected Mann-Whitney *U* test was applied, and the p-value was set at 0.017.
BHS = Beck Hopelessness Scale; CG = Control Group; DG = Depressed Group; HARS = Hamilton Anxiety Rating Scale; HDRS = Hamilton Depression Rating Scale; NTSG = Near-Term Suicidal Group.
† Comparison between CG and DG.
‡ Comparison between CG and NTSG.
§ Comparison between DG and NTSG.

Details of the model metrics, presented as mean values from the 10-fold cross-validation, are shown in Table 1.

Among the analyzed speech features, the MFCC feature representation was found to be more successful for the tasks "high suicide risk or not" as well as "depression or high suicide risk." The accuracy, precision, sensitivity, F1 score, and specificity for the "high suicide risk or not" task using MFCC features were 0.900, 0.883, 0.933, 0.904, and 0.866, respectively. For the "depression or high suicide risk" task, MFCC features yielded accuracy, precision, sensitivity, F1 score, and specificity values of 0.683, 0.662, 0.767, 0.700, and 0.600, respectively. For the task "depression or not" VGGish feature representation was found to be more successful; accordingly, the accuracy, precision, sensitivity, F1 score, and specificity were 0.733, 0.810, 0.700, 0.716, and 0.767, respectively.

Discussion

Accurate and scalable tools for suicide risk detection remain a pressing need in clinical psychiatry, where assessment often relies on subjective self-report. This study evaluated whether voice-derived features, analyzed via machine learning, can serve as objective biomarkers for identifying depression and near-term suicide risk. Among the extracted features, MFCCs and deep features (VGGish) were particularly effective. MFCCs yielded the highest classification performance for identifying individuals at imminent suicide risk (accuracy: 90%), while VGGish features better differentiated depressed individuals from healthy controls (accuracy: 73.3%). These findings offer preliminary evidence that subtle variations in spectral and learned deep voice features may encode clinically meaningful distinctions across the suicidality spectrum.

The development of accurate and scalable tools for suicide risk assessment remains a critical priority in clinical psychiatry, where diagnostic evaluations often depend on subjective self-report. We investigated the potential of voice-derived features as objective biomarkers for detecting depression and acute suicide risk using machine learning-based classification. Among the extracted features, MFCCs exhibited the highest performance in identifying individuals at imminent suicide risk (accuracy: 90%) and in distinguishing between depressed and high-risk individuals (accuracy: 68.3%). In contrast, VGGish demonstrated superior accuracy in distinguishing depressed individuals from healthy controls (accuracy: 73.3%). These results provide preliminary evidence that acoustic and deep learned voice features encode clinically salient differences across the suicidality continuum.

We observed that individuals in the NTSG had significantly higher HDRS scores than those in the non-suicidal DG, which is consistent with evidence linking greater depression severity to increased suicide risk.[2] Rather than conceptualizing this difference as a confounding variable, we interpret it within a dimensional framework in which suicidality is understood as a continuum of risk. However, some individuals in the high-risk group had lower HDRS scores than participants in the non-suicidal group, which highlights clinical heterogeneity and suggests that vocal features may reflect suicide risk beyond depression severity. This aligns with studies suggesting that acoustic biomarkers capture shifts in emotional regulation, hopelessness, and impulsivity.[13,39] Although we did not control for depression severity, the HDRS scores provide contextual insight. Future studies should apply models that adjust for symptom severity to better isolate the contribution of voice. Moreover, by training separate classification models for each group, our study enabled the identification of risk-specific acoustic signatures that are tailored to distinct clinical decision-making scenarios, such as differentiating suicidality from depression, and contribute to psychiatric evaluation. The higher BHS scores in the control group may be related to social stressors and economic difficulties.

VGGish, a deep audio representation model trained on the AudioSet dataset,[40] could effectively distinguish depressed individuals from controls in our study. Furthermore, in our other two classification tasks, VGGish emerged as the second most effective feature after MFCC, showing promise in identifying individuals at high suicide risk. While VGGish has previously been used in fields such as dementia screening,[41] respiratory disease classification,[42] and COVID-19 detection,[43,44] to our knowledge, this is the first study that has used it to classify suicide risk levels. The success of the VGGish model in our study may be attributed to several key factors. First, its training on a large and diverse dataset likely enhanced its generalizability to new speech data. Second, the use of log-mel spectrograms, which robustly represent the spectral characteristics of speech signals, contributed to the model's effective feature extraction. Finally, the deep learning architecture of VGGish enabled it to capture complex patterns and relationships, allowing the detection of subtle and clinically meaningful voice features associated with psychiatric states.

MFCCs with the most widely used spectral features reflect vocal tract behavior and muscle tension by capturing the frequency distribution of speech using mel filters modeled after human cochlear response.[13,33,45,46] MFCCs have previously been used to classify various psychiatric conditions, including depression and suicidal ideation.[13,20,23,47,48] In our study, MFCC features achieved the highest classification performance in both the "high suicide risk or not" and "depression or high suicide risk" tasks. These findings indicate that MFCCs not only distinguish individuals at high suicide risk from healthy controls but also from depressed individuals with lower suicide risk. The high classification accuracy indicates that MFCCs may detect vocal signatures specifically associated with suicide risk, potentially reflecting dimensions distinct from depression. The discriminative power of MFCCs in assessing depression and suicide risk is consistent with previous research. For example, Taguchi et al.[48] reported significantly higher MFCC2 levels in 36 patients with depression than 36 healthy controls, independent of age and gender, yielding a sensitivity of 77.8% and specificity of 86.1% in distinguishing the groups. Similarly, Ozdas et al.[49] reported that MFCCs could distinguish between healthy individuals, depressed patients, and those with high suicide risk with 75-80% accuracy. Keskinpala et al.[19] found that MFCCs achieved strong classification performance during reading tasks, reaching 80.62% accuracy in men and 73.37% in women. More recently, Min et al.[21] analyzed 348 voice recordings from 104 individuals with mood disorders and achieved 79% accuracy in predicting within-subject suicide risk. The current study also demonstrated that many MFCC parameters were significantly associated with suicide risk. These findings provide robust evidence supporting the use of MFCCs as reliable biomarkers for the objective assessment of depression and suicide risk. Consistent with prior literature, our findings show that MFCCs most effectively distinguished the NTSG from both CG and DG, indicating their potential to capture suicide-specific vocal signatures relevant to clinical assessment.

Speech-based machine learning techniques offer a scalable, objective method for psychiatric assessment by enabling non-invasive access to emotional and cognitive states. Supporting prior evidence, our findings show that MFCCs performed best in detecting high suicide risk, while VGGish effectively distinguished depression from healthy states. These results underscore the potential utility of voice-based tools in a variety of clinical and technological contexts, particularly where conventional evaluations are limited by time, stigma, or lack of access. The feasibility of using voice biomarkers in telepsychiatry or mobile apps opens new possibilities for early detection and monitoring. As global demand for accessible mental health solutions increases, voice analysis stands out as a practical and interpretable tool to support diagnosis and suicide prevention.

The clinical utility of speech analysis is grounded in its neurobiological underpinnings. Speech production is a neuromotor process involving cortical, subcortical, and peripheral systems that regulate mood and cognition, functions that are often disrupted in depression and suicidality. Speech alterations, such as reduced prosodic variation, spectral flattening, or slowed articulation, may reflect disruptions in these neural pathways.[13] Studies have linked depression to measurable phonation and articulation changes, likely due to psychomotor slowing and autonomic dysregulation.[13,48] Supporting this, affective disturbances such as depression and suicidality were reported to alter the somatic and autonomic nervous systems, thereby impacting the muscles responsible for phonation and articulation.[49] Features like MFCC can capture these subtle vocal tract changes. Thus, voice-based models are not merely statistical classifiers but capture biologically meaningful signals, supporting their use as objective markers of psychiatric status.[50]

While this study offers promising insights into voice-based detection of suicide risk, several methodological limitations warrant consideration. First, the small sample size constrains the generalizability of the findings and calls for replication in larger cohorts. Second, although age differences between groups may have influenced vocal feature interpretation, prior findings[48] indicate that key MFCC parameters, such as MFCC2, are age-independent. Nevertheless, potential age-related effects should be evaluated using statistical tools. Third, depression severity was not included as a covariate in our models, which may have influenced the classification outcomes, especially given the close association between depressive symptomatology and suicidality. Additionally, clinical heterogeneity remains a concern. Although participants with comorbid psychiatric diagnoses and psychotic features were excluded, the study groups used different medications. According to subgroup analyses, medication status did not have statistically significant effects; however, future research involving larger and pharmacologically homogeneous samples would be better suited to isolate and identify such influences. Finally, the use of cross-sectional data precludes any conclusions about the temporal stability or predictive value of voice-based features. Longitudinal designs will be essential to determine whether vocal biomarkers can prospectively track or anticipate shifts in suicide risk over time.

In conclusion, the current study offers preliminary but clinically relevant evidence that voice-derived acoustic and deep features can function as biomarkers for distinguishing healthy individuals, depressed patients, and those at near-term suicide risk. To our knowledge, this is the first study to apply the VGGish deep audio embedding model to classify suicide risk levels in a clinical population, marking a novel contribution to the growing field of speech-based psychiatric assessment. The high classification performance of both MFCC and VGGish features supports the potential of voice analysis as an objective, scalable, and non-invasive tool in mental health care. This approach holds particular promise in settings with limited access to psychiatric services or where traditional evaluations are constrained by time, stigma, or communication barriers. Nevertheless, limitations such as the relatively small sample size and clinical heterogeneity should be addressed in future studies. Such large, demographically balanced, and longitudinally designed research will be crucial to enhancing the model's robustness and clinical utility. Integrating speech-based models into standard psychiatric evaluation could help enable earlier detection, more accurate risk stratification, and improved suicide prevention strategies.

Disclosure

The authors report no conflicts of interest.

Data availability statement

The data that support this study may be available from the authors upon reasonable request.

Author contributions

SO: Writing – review & editing, Writing – original draft, Conceptualization, Methodology, Investigation, Formal Analysis, Visualization, Data-curation.
MA: Writing – review & editing, Supervision, Methodology, Conceptualization, Investigation, Project Administration, Funding Acquisition.
MS: Writing – review & editing, Methodology, Software, Formal Analysis, Validation, Conceptualization, Visualization.
ŞG: Writing – review & editing, Methodology, Investigation, Conceptualization, Formal Analysis.
OÇ: Writing – review & editing, Methodology, Resources, Supervision, Validation.
IUB: Writing – review & editing, Investigation, Data-curation.
AYA: Writing – review & editing, Investigation, Methodology, Supervision, Resources.

All authors have read and approved of the final version to be published.

Handling Editor: Rodolfo Damiano

References

1 World Health Organization (WHO). Suicide worldwide in 2019: global health estimates. New York: World Health Organization; 2021.

2 Hawton K, Casañas I Comabella C, Haw C, Saunders K. Risk factors for suicide in individuals with depression: a systematic review J Affect Disord. 2013;147:17-28.

3 Mann JJ. Neurobiology of suicidal behaviour. Nat Rev Neurosci. 2003;4:819-28.

4 Briley M, Lépine. The increasing burden of depression. Neuropsychiatr Dis Treat. 2011;7:3-7.

5 Turecki G, Brent DA. Suicide and suicidal behaviour. Lancet. 2016;387:1227-39.

6 García-Gutiérrez MS, Navarrete F, Sala F, Gasparyan A, Austrich-Olivares A, Manzanares J. Biomarkers in psychiatry: concept, definition, types and relevance to the clinical reality. Front Psychiatry. 2020;11:432.

7 Lozupone M, La Montagna M, D'Urso F, Daniele A, Greco A, Seripa D, et al. The role of biomarkers in psychiatry. In: Guest PC, editor. Reviews on biomarker studies in psychiatric and neurodegenerative disorders. Cham: Springer International Publishing; 2019. p. 135-62.

8 Johnston JN, Campbell D, Caruncho HJ, Henter ID, Ballard ED, Zarate CA. Suicide biomarkers to predict risk, classify diagnostic subtypes, and identify novel therapeutic targets: 5 years of promising research. Int J Neuropsychopharmacol. 2022;25:197-214.

9 Luoma JB, Martin CE, Pearson JL. Contact with mental health and primary care providers before suicide: a review of the evidence. Am J Psychiatry. 2002;159:909-16.

10 Stene-Larsen K, Reneflot A. Contact with primary and mental health care prior to suicide: a systematic review of the literature from 2000 to 2017. Scand J Public Health. 2019;47:9-17.

11 Obegi JH. How common is recent denial of suicidal ideation among ideators, attempters, and suicide decedents? A literature review. Gen Hosp Psychiatry. 2021;72:92-5.

12 Corcoran CM, Mittal VA, Bearden CE, Gur RE, Hitczenko K, Bilgrami Z, et al. Language as a biomarker for psychosis: a natural language processing approach. Schizophr Res. 2020;226:158-66.

13 Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 2015;71:10-49.

14 Kraepelin E. Manic-depressive insanity and paranoia. Edimburg: E. & S. Livingstone; 1921.

15 Cohen AS, Rodriguez Z, Opler M, Kirkpatrick B, Milanovic S, Piacentino D, et al. Evaluating speech latencies during structured psychiatric interviews as an automated objective measure of psychomotor slowing. Psychiatry Res. 2024;340:116104.

16 Iyer R, Meyer D. Detection of suicide risk using vocal characteristics: systematic review. JMIR Biomed Eng. 2022;7:e42386.

17 Silverman SE. Vocal parameters as predictors of near-term suicidal risk, U.S. Patent. 1992;5:148.

18 France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes DM. Acoustical properties of speech as indicators of depression and suicidal risk. I IEEE Trans Biomed Eng. 2000;47:829-37.

19 Keskinpala HK, Yingthawornsuk T, Wilkes DM, Shiavi RG, Salomon RM. Screening for high risk suicidal states using mel-cepstral coefficients and energy in frequency bands. In: 15th European Signal Processing Conference; 2007; Poznan, Poland. p. 2229-33.

20 Belouali A, Gupta S, Sourirajan V, Yu J, Allen N, Alaoui A, et al. Acoustic and language analysis of speech for suicidal ideation among US veterans. BioData Min. 2021;14:11.

21 Min S, Shin D, Rhee SJ, Park CHK, Yang JH, Song Y, et al. Acoustic analysis of speech for screening for suicide risk: machine learning classifiers for between-and within-person evaluation of suicidality. J Med Internet Res. 2023;25:e45456.

22 Yingthawornsuk T, Keskinpala HK, Wilkes DM, Shiavi RG, Salomon RM. Direct acoustic feature using iterative EM algorithm and spectral energy for classifying suicidal speech. In: 8th Annual Conference of the International Speech Communication Association; August 27-31, 2007; Antwerp, Belgium.

23 Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig Otolaryngol. 2020;5:96-116.

24 Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960;23:56-62.

25 Hamilton M. Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol. 1967;6:278-96.

26 Akdemir A, Dönbak Örsel Sibel, Dağ İhsan, Türkçapar M. Hakan, İşcan Nalan, Özbay Haluk. Hamilton depresyon derecelendirme ölçeği (HDDÖ)'nin geçerliliği- güvenirliliği ve klinikte kullanımı. Psikiyatri Psikoloji Psikofarmakol Derg. 1996;4:251-9.

27 Hamilton M. The assessment of anxiety states by rating. Br J Med Psychol. 1959;32:50-5.

28 Yazıcı MK, Demir B, Tanrıverdi N, Karaağaoğlu E, Yolaç P. Hamilton anksiyete değerlendirme öiçeği, değerlendiriciler arası güvenirlik ve geçerlik çalışması. Turk Psikiyatri Derg. 1998;9:114-7.

29 Beck AT, Weissman A, Lester D, Trexler L. The measurement of pessimism: the hopelessness scale. J Consult Clin Psychol. 1974; 42:861-5.

30 Seber G, DiLbaz N, Kaptanoğlu C, TekiN D. Umutsuzluk Ölçeği: Geçerlilik ve Güvenilirliği. Kriz Derg. 1998;001-4.

31 Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry. 1978;133:429-35.

32 Karadağ F, Oral T, Yalçın FA, Erten E. [Reliability and validity of Turkish translation of Young Mania Rating Scale]. Turk Psikiyatri Derg. 2002;13:107-14.

33 Erdem ES, Sert M. Efficient recognition of human emotional states from audio signals. In: 2014 IEEE International Symposium on Multimedia; 10-12 Dec 2014; NW Washington, DC, United States. p. 139-42.

34 Wang J, Zhang L, Liu T, Pan W, Hu B, Zhu T. Acoustic differences between healthy and depressed people: a cross-situation study. BMC Psychiatry. 2019;19:300.

35 Sarman S, Sert M. Audio based violent scene classification using ensemble learning. In: 2018 6th International Symposium on Digital Forensic and Security (ISDFS); 2018; Antalya, Turkey. p. 1-5.

36 Sert M, Baykal B, Yazici A. Generating expressive summaries for speech and musical audio using self-similarity clues. In: 2006 IEEE International Conference on Multimedia and Expo; 2006; Toronto, Canada. p. 941-4.

37 Hershey S, Chaudhuri S, Ellis DPW, Gemmeke JF, Jansen A, Moore RC, et al. CNN architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA; 2017. p. 131-5.

38 Selbes B, Sert M. Multimodal vehicle type classification using convolutional neural network and statistical representations of MFCC. In: 4th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Lecce, Italy; 2017. p. 1-6.

39 Alghowinem S, Zhang X, Breazeal C, Park HW. Multimodal region-based behavioral modeling for suicide risk screening. Front Comput Sci. 2023;5:990426.

40 Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC, et al. Audio Set: an ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA; 2017. p. 776-80.

41 Chlasta K, Wołk K. Towards computer-based automated screening of dementia through spontaneous speech. Front Psychol. 2021;11: 623237.

42 Choi Y, Lee H. Interpretation of lung disease classification with light attention connected module. Biomed Signal Process Control. 2023; 84:104695.

43 Acar YA, Sert M, Ozen D, Bayar IU, Bayar MM, Kavakli M, et al. Early detection of respiratory diseases caused by COVID-19 and integration into tele-health service via speech, voice and cough analysis software. 2021 [cited 2025 Nov 21]. baskent.edu.tr/∼msert/research/pubs/covid19-best-paper-abs-0217.pdf

44 Despotovic V, Ismael M, Cornil M, Call RM, Fagherazzi G. Detection of COVID-19 from voice, cough and breathing patterns: dataset and preliminary results. Comput Biol Med. 2021;138:104944.

45 Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.

In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4; Illinois, USA; Aug 1980. p. 357-66.
46 Kucukbay SE, Sert M. Audio-based event detection in office live environments using optimized MFCC-SVM approach. In: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015). Anaheim, CA, USA; 2015. p. 475-80.
47 Akkaralaertsest T, Yingthawornsuk T. Comparative analysis of vocal characteristics in speakers with depression and high-risk suicide. Int J Comput Theoty Eng. 2015;7:448-52.
48 Taguchi T, Tachikawa H, Nemoto K, Suzuki M, Nagano T, Tachibana R, et al. Major depressive disorder discrimination using vocal acoustic features. J Affect Disord. 2018;225:214-20.
49 Ozdas A, Shiavi RG, Wilkes DM, Silverman MK, Silverman SE. Analysis of vocal tract characteristics for near-term suicidal risk assessment. Methods Inf Med. 2004;43:36-8.
50 Gideon J, Provost EM, McInnis M. Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Shanghai, China; 2016. p. 2359-63.