

Article

Addressing the Dissimulation Problem in the Assessment of Suicide Risk Using CNN-Based Voice Analysis: A Proof-of-Concept Study

Adela Magdalena Ciobanu ^{1,2} , Ana Voichita Tebeanu ^{3,*}, Vlad Tau ^{2,4,*}, Anca Ana Chendea ^{1,2}, Eduard Dan Franti ^{5,6}, Catalin Niculae ⁷, Claudiu Ionut Vasile ^{8,9} , George Florian Macarie ^{3,10}, Liliana Neagu ¹, Monica Dascalu ¹¹, Cristian Ioan Stoica ¹, Marius Moga ¹  and Gabriela Iorgulescu ¹

¹ Faculty of Medicine, University of Medicine and Pharmacy “Carol Davila”, 020021 Bucharest, Romania; adela.ciobanu@umfcd.ro (A.M.C.); ana.talasman@umfcd.ro (A.A.C.); liliana.neagu@umfcd.ro (L.N.); cristian.stoica@umfcd.ro (C.I.S.); marius.moga@umfcd.ro (M.M.); gabriela.iorgulescu@umfcd.ro (G.I.)

² “Prof. Dr. Al. Obregia” Psychiatry Clinical Hospital, 041914 Bucharest, Romania

³ Department for Teachers’ Training and Social Sciences, National University of Science and Technology Politehnica Bucharest, 060042 Bucharest, Romania; george.macarie@upb.ro

⁴ Eden Health SRL, 012256 Bucharest, Romania

⁵ Microsystems in Biomedical and Environmental Applications Laboratory, IMT Bucharest, 077190 Voluntari, Romania; eduard.franti@imt.ro

⁶ Research Institute for Artificial Intelligence “Mihai Draganescu”, 050711 Bucharest, Romania

⁷ Areus Technology, 022142 Bucharest, Romania; catalin.niculae@areusdev.com

⁸ “Elisabeta Doamna” Psychiatry Hospital, 800179 Galati, Romania; ionut.vasile@ugal.ro

⁹ Faculty of Medicine and Pharmacy, “Dunarea de Jos” University, 800010 Galati, Romania

¹⁰ Clinica PSIHOGER, 700126 Iasi, Romania

¹¹ Faculty of Electronics, Telecommunications and Information Technology, Department of Electronic Devices, Circuits and Equipment, National University of Science and Technology Politehnica Bucharest, 060042 Bucharest, Romania; monica.dascalu@upb.ro

* Correspondence: voichita.tebeanu@upb.ro (A.V.T.); contact@vladtau.com (V.T.)

Abstract

Background/Objectives: Suicide remains a major public health concern, with over 700,000 deaths annually worldwide. Current risk assessment is limited by the tendency of at-risk patients to deny suicidal ideation, with denial rates of approximately 50% among ideators and up to 78% among inpatient suicide decedents. **Methods:** This proof-of-concept study investigated whether deep learning applied to patient speech could distinguish psychiatric inpatients admitted following a recent suicide attempt from patients with severe depression without documented suicidal ideation. A multi-scale two-dimensional convolutional neural network with residual connections and spatial attention (Multi-Scale CNN) was trained on mel-spectrograms generated exclusively from patient speech extracted by automatic speaker diarization followed by manual verification. The dataset comprised recordings from 88 psychiatric inpatients (59 suicide attempters and 29 patients with severe depression without suicidal ideation). Model performance was evaluated using repeated patient-level stratified five-fold cross-validation, ensuring complete separation of participants between folds. Additional female-only, exploratory male-only, and repeated sex-matched analyses were performed to evaluate the potential influence of sex imbalance. **Results:** Repeated patient-level five-fold cross-validation yielded a mean classification accuracy of $72.7 \pm 6.4\%$ with a mean ROC-AUC of 0.824 ± 0.054 , a sensitivity of $88.3 \pm 9.5\%$, and a specificity of $41.3 \pm 25.0\%$. Female-only analysis maintained good discriminative performance (ROC-AUC 0.866 ± 0.123), whereas repeated sex-matched analyses produced lower performance (ROC-AUC 0.621 ± 0.156), indicating that sex imbalance contributed to, but did not fully explain, the observed discrimination. The male-only analysis was considered exploratory because of the limited number of male patients in the depression



Academic Editor: Motohiro Okada

Received: 4 June 2026

Revised: 24 July 2026

Accepted: 30 July 2026

Published: 4 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

group. Conclusions: These findings support the feasibility of voice-based patient-level analysis as a complementary objective approach for suicide risk assessment. Although the study remains exploratory because of the limited sample size and the absence of external validation, the revised validation strategy and sex-controlled analyses provide a substantially more robust estimate of model performance. Further multicenter prospective studies are required before clinical implementation.

Keywords: suicidal behavior; AI algorithms; speech emotion detection; convolutional neural network; multi-scale CNN; mel-spectrogram; voice biomarker; dissimulation

1. Introduction

Suicide and suicidal behavior remain major public health concerns, with measurable impact on mortality and service demand. The World Health Organization estimates that around 727,000 people die by suicide each year worldwide [1]. In the European Union, Eurostat reported 49,042 deaths by suicide in 2022, and among people aged 15–29, roughly one in six deaths was due to suicide [2]. In the United States, the Centers for Disease Control and Prevention reported 49,316 suicide deaths in 2023 [3].

Clinical risk assessment continues to face a basic difficulty: suicidal crises are heterogeneous and, for rare outcomes, even strong risk factors translate poorly into individual-level prediction. A landmark meta-analysis of 365 studies spanning 50 years of research found that prediction of suicidal thoughts and behaviors was only slightly better than chance for all outcomes, and that the accuracy of prediction had not improved appreciably over five decades [4]. Meta-analytic evidence suggests that about one in five people who make an index suicide attempt will reattempt, with risk accumulating over time and varying by clinical and demographic factors [5]. The WHO also emphasizes the role of proximal drivers such as mental disorders, harmful use of alcohol, access to lethal means, and social stressors [1].

Several reviews have shown that commonly used clinical instruments and risk scales have limited positive predictive value for future suicidal behavior, especially when applied outside narrow research contexts [6]. A systematic review and simulation of prediction models for suicide attempts and deaths demonstrated that even models with ostensibly good discrimination (high AUC) yield positive predictive values near zero when applied to low-prevalence populations, underscoring the gap between statistical performance and clinical utility [7]. This limitation has motivated a shift toward data-driven approaches that can integrate higher-dimensional information, including clinical history, service use, and behavioral signals.

Recent appraisal work highlights that the field has produced many prediction models for self-harm and suicide, but few have been externally validated, and many development studies show high risk of bias due to sample size, handling of missing data, and incomplete evaluation of calibration or overfitting [8]. These concerns are echoed in systematic reviews focused on machine learning (ML) applications in psychiatric populations, where study design and reporting often limit clinical translation [9]. A recent meta-analysis also shows that favorable global metrics can coexist with low positive predictive value in low-prevalence settings, which constrains how risk flags can be used in practice [10].

These empirical limitations have also prompted a broader debate regarding the purpose and conceptual foundations of suicide risk assessment. Teismann et al. argue that the basic structure of clinical risk assessment may remain useful, provided that its purpose is recalibrated away from precise prediction and toward guiding clinical understanding,

safety planning, and therapeutic intervention [11]. A more radical position is advanced by Large and Ryan, who question the utility of conventional risk categorization and propose that clinicians should instead evaluate whether patients are currently capable of making autonomous decisions regarding their own welfare [12]. At a more fundamental level, Baston argues that the concept of “risk” itself requires reconsideration, suggesting a move beyond prediction toward a framework that better captures the dynamic, contextual, and relational nature of suicidal crises [13]. The present study does not seek to resolve these conceptual debates or to replace comprehensive clinical assessments; rather, it examines whether voice-derived signals may provide complementary information when conventional assessment is limited by incomplete or deliberately withheld self-report.

A fundamental limitation shared by all self-report-based screening instruments is their vulnerability to patient dissimulation. A narrative review examining the empirical basis for denial of suicidal ideation found that approximately 50% of individuals with suicidal ideation deny it during clinical interviews, and approximately half of suicide decedents denied suicidal ideation in the weeks or months preceding their deaths [14]. In a study of 157 patients who died by suicide while receiving care, two-thirds had denied suicidal ideation when last asked by a clinical practitioner, and half of those who denied ideation were dead within two days [15]. A chart review of 76 inpatient suicides found that 78% of patients had denied suicidal ideation at their last communication before death [16]. A psychological autopsy study in Norway reported that 80% of informants perceived some form of dissembling or masking behavior in the deceased, with older decedents exhibiting more dissimulation [17]. These findings indicate that reliance on patient disclosure as a gateway to suicide risk assessment is fundamentally flawed, and they motivate the search for objective biomarkers that capture risk-relevant signals below the threshold of voluntary control.

The period immediately following discharge from psychiatric hospitalization represents a window of extremely elevated suicide risk. A meta-analysis of 100 studies comprising 17,857 suicides found that the pooled post-discharge suicide rate was 484 per 100,000 person-years overall, rising to 1132 per 100,000 person-years within the first three months, roughly 100 times the global baseline rate [18]. More granular analyses show that risk is highest in the first days: a Finnish population-based study of 193,197 hospitalizations for depression reported incidence rates of 6062 per 100,000 person-years on days 0–3 and 3884 per 100,000 on days 4–7 after discharge [19]. A separate meta-analysis estimated pooled suicide rates of 2950 per 100,000 person-years in the first week and 2060 per 100,000 in the first month post-discharge [20]. Across studies, approximately 26% of post-discharge suicidal events occur within the first month, and 41% within three months [21]. These data establish the post-discharge period as a critical window where continuous, non-self-report monitoring could have the greatest clinical impact.

Parallel to electronic health record modeling, there has been growing interest in speech and language as potentially informative, low-burden signals associated with suicide risk. Research on voice stress analysis has established that the autonomic nervous system modulates acoustic parameters including fundamental frequency, voice quality, and spectral energy distribution in ways that are partially independent of volitional control [22]. A comprehensive review of speech analysis for depression and suicide risk assessment documented recurring associations between clinical status and acoustic measures, while identifying methodological challenges including small samples, inconsistent feature sets, and limited external validation [23]. More recent systematic reviews have reported associations between risk status and features such as F0 statistics, jitter, and MFCCs, while emphasizing variability in datasets, elicitation tasks, and evaluation protocols [24]. In prospective work, acoustic patterns have been examined both between and within

individuals, suggesting that person-specific baselines can matter when interpreting vocal change [25]. Other studies have combined paralinguistic features with deep learning for screening in specific contexts, including crisis hotline calls [26] and mobile-phone speech diaries, where cross-dataset evaluation highlights distribution shift as a barrier to deployment [27].

State of the Art

AI-supported suicide risk assessment is currently dominated by supervised prediction models trained on electronic health records (EHR), registry data, and structured clinical information. Reviews spanning the last decade converge on two points: discrimination can be moderate to high in some settings, but estimates are difficult to compare because outcomes, sampling strategies, and validation practices differ. Lejeune et al. reviewed 17 studies published between 2014 and 2020 and reported AUC (area under the curve) values ranging from 0.604 to 0.947, with most work relying on retrospective designs and routinely collected clinical data [28]. Kusuma et al. synthesized longitudinal predictions of ideation, attempts, and deaths, reporting a pooled AUC of 0.86 with pooled sensitivity of 0.66 and specificity of 0.87, while also finding that risk of bias remained common across included studies [29]. In a systematic review and meta-analysis restricted to clinical studies (excluding natural language processing and image-based approaches), Ehtemam et al. pooled 41 studies and estimated overall accuracy of 0.78 and AUC of 0.77 (sensitivity 0.69, specificity 0.81), with substantial heterogeneity [30].

The most recent broad appraisal of self-harm and suicide prediction models similarly found persistent issues in methodological rigor and transparency, with a small fraction of models undergoing external validation and many studies rated at high risk of bias under PROBAST criteria (the Prediction model Risk Of Bias ASsessment Tool) [8]. Pigoni et al. also emphasized that calibration is rarely assessed and that reporting biases likely inflate some performance estimates [9].

Beyond EHR and administrative data, recent reviews map a diversification of data sources used for AI-enabled prevention, including social media content, crisis service transcripts, ecological momentary assessment, sensor-derived digital phenotyping, and multimodal combinations. Randomized evidence on whether AI-enabled tools improve suicide-related outcomes is still limited. Fernandez-Quijano et al. reviewed randomized controlled trials published through 31 May 2025 and identified six trials ($n = 793$) spanning three functional domains: machine-learning risk prediction, fully automated digital interventions, and AI-assisted treatment allocation. Across the risk-prediction trials, short-term outcomes were stratified with modest discrimination (reported AUC values around 0.70 and accuracies up to 0.67). The two fully automated interventions focused on service processes, including reductions in counsellor response latency and increased crisis-service uptake (reported as a 23% increase in uptake). The treatment-allocation trial reported fewer suicidal events when randomization aligned with the model recommendation. Despite generally good methodological quality (median PEDro score 8/10), the authors rated certainty of evidence as low because of small samples and imprecision, and they called for larger multisite trials with longer follow-up and reporting aligned with CONSORT-AI, including equity-focused evaluation [31].

Work focused on adolescents is expanding, partly because youth suicide and self-harm present distinct developmental, social, and clinical patterns. A 2025 systematic review and meta-analysis on ML prediction of suicide-related outcomes in adolescents reported moderate pooled performance and highlighted that small samples, imbalanced outcomes, and limited prospective validation remain common [32].

Passive sensing and digital phenotyping represent another active strand of research, aiming to infer near-term risk from smartphones and wearable devices. In a systematic review with database searches through 24 June 2024, Buscher et al. identified 11 prediction studies, 10 feasibility studies, and 7 study protocols focused on passive signals related to suicidal thoughts and behaviors. Across the prediction studies, model performance was generally lower for passive compared with active inputs, and in three of four studies that tested incremental value, passive data did not improve prediction beyond active data. The authors also reported substantial methodological and reporting shortcomings based on PROBAST, including heterogeneous outcomes and limited validation, while feasibility findings suggested that passive monitoring can be acceptable in high-risk clinical samples, with adherence varying by device and study context [33]. Complementary work has addressed the more common reality of irregular and incomplete active self-report time series. Using questionnaire data collected in the EMMA (Ecological Mental Momentary Assessment) smartphone platform from 89 at-risk participants, Quéllec et al. modeled suicidal ideation as a multivariate time series with substantial missingness and used a transformer-based imputation approach (a modified SAITS variant). They reported discrimination between high and low ideation on the same day (AUC = 0.804) and three days ahead (AUC = 0.769). Beyond prior ideation, permutation-importance analyses highlighted items related to psychological pain, agitation, emotional tension, well-being, and protective factors (for example, contact with relatives and leisure activities) as contributors to model performance [34].

Natural language processing (NLP) is increasingly applied to crisis lines and digital support services, where text carries both risk information and conversational context. A recent *Frontiers in Medicine* study analyzed real-time crisis chats using a theory-driven lexicon and interpretable modeling to examine how risk-related constructs evolve across conversation stages, reporting differences by gender and age strata [35]. These designs align with a broader shift toward models that can be audited and interpreted by clinicians and service providers, especially in settings where decisions are time sensitive.

Speech-based modeling remains a smaller, active area. The literature spans clinical interviews, helpline calls, and web-collected recordings, with features ranging from classical prosody and voice quality measures to self-supervised embeddings. Min et al. reported associations between suicide risk and acoustic patterns from both between- and within-person perspectives, suggesting that person-specific baselines matter when interpreting vocal change [25]. In crisis-line settings, Amiriparian et al. applied convolutional architectures to call audio to classify callers as at-risk or not at-risk and noted that transfer across call centers and languages is an open challenge, though the small sample size (20 patients) limits general claims [36]. Krautz et al. extended speech-based prediction to suicidal behavior outcomes and included external validation across countries but still reported performance drops under cross-national transfer [37]. The authors claimed that this was the first study to successfully predict actual suicidal behavior rather than surrogate markers, marking a major step forward in suicide prevention. Cross-dataset experiments with mobile-phone speech diaries similarly report strong sensitivity to dataset shift [27]. Recent work using crisis hotline calls shows that acoustic features can support risk classification, but the low base rate of suicide constrains positive predictive value and makes operational thresholding consequential [26].

Related paralinguistic work has used CNN-based pipelines on time-frequency speech representations for Romanian-language emotion recognition and interactive applications, including companion-robot emotion classification and voice-based intention detection [38,39]. This aligns with broader speech emotion recognition literature, where

convolutional models trained on spectrogram-like representations remain a common design choice [40,41].

Evidence from a Romanian speech dataset suggests that negative arousal can persist for minutes after an emotionally triggering interaction and may intensify as a stressful event is anticipated, motivating models that incorporate temporal context rather than isolated clips [42].

Controlled experiments have also used CNN-based speech analysis to quantify induced negative affect and irritability from short segments [43].

Transfer across recording contexts and populations remains difficult; nevertheless, cross-cultural evaluation has been explored in other non-semantic vocalization tasks, such as infant vocalization recognition tested across Australian and Romanian recordings [44,45].

Benchmark datasets and shared evaluation frameworks are beginning to appear, addressing a long-standing fragmentation problem in speech and multimodal research. The SpeechWellness challenge released a speech dataset collected from adolescents and reported baseline systems intended to support comparable evaluation across teams [46].

In parallel, work on network-based feature representations and other structured acoustic modeling has been proposed to improve robustness and interpretability when generalizing across tasks and speakers [47]. Across modalities, the current evidence base suggests that progress depends less on novel architecture alone and more on data quality, reliable labels, transparent reporting, and external validation in settings close to intended use [8,9,24,31].

2. Materials and Methods

The objective of this proof-of-concept study was to investigate whether deep learning applied to voice recordings can identify acoustic patterns associated with recent suicidal behavior and distinguish patients with suicide attempts from patients with depression without recent suicidal behavior. This hypothesis is based on the following considerations, detailed below:

1. We hypothesized that individuals in a pre-suicidal state experience, before taking the final action by which they will take their own life, very intense emotions specific to suicide, due to the profound ambivalence and psychological turmoil that characterize the pre-suicidal state.
2. We hypothesized that intense emotional states influence voice production through neurophysiological mechanisms and that part of these acoustic alterations may remain detectable beyond the immediate emotional episode, either because of residual autonomic activation or because they reflect a more persistent post-crisis physiological state.
3. We hypothesized that acoustic patterns associated with suicidal behavior have a different signature than other emotions in normal life situations, that they are intense, that they remain embedded in the voice for extended periods, and that they can be detected with AI algorithms.

Importantly, the proposed approach is not intended to replace clinical evaluation. Rather, it investigates whether voice analysis may provide an additional objective source of information that could complement conventional suicide risk assessment, particularly in situations where patients deny suicidal ideation.

2.1. Methodology Rationale

Comparing voice recordings from patients with suicide attempts to those from patients with depression (without recent suicidal behavior) is methodologically defensible because depression functions as an active clinical control: it shares many non-specific determinants

of voice (e.g., psychomotor slowing, reduced affective reactivity, sleep disturbance, medication exposure, and inpatient/clinical interaction effects), so any residual separability is more plausibly attributable to suicide-specific state markers rather than “depression in general.”

This active-control logic has clear precedent in suicide-speech research: foundational acoustic work explicitly contrasted high-risk suicidal patients with major depressed and non-depressed groups, showing that formant/power-distribution measures and related spectral features [48] as well as lower-order MFCCs [49] and excitation/phonation features such as jitter and glottal spectral slope [50] can discriminate suicidal patients not only from healthy controls but also from depressed comparators.

More recent studies continue to operationalize depression (or mood-disorder cohorts) as the key comparator when developing ML models for suicide risk from speech, including interview-based acoustic prediction models within mood-disorder samples [25].

Complementary evidence from read-speech and naturalistic recordings indicates that suicidal ideation/attempt status is detectable from voice quality, prosodic/phonatory, and spectral markers [49–52], while systematic synthesis emphasizes that demonstrating incremental validity requires comparisons against clinically similar cohorts (often including depressed patients), not only healthy controls [23].

Although depression was selected as an active clinical comparator to minimize non-specific vocal differences, the two clinical groups differed in sex distribution. Because sex influences several acoustic characteristics of speech, additional female-only, exploratory male-only, and repeated sex-matched analyses were performed to evaluate whether the observed classification performance remained after partially controlling for this potential source of confounding. The primary performance estimates reported in this study are those obtained from repeated patient-level stratified five-fold cross-validation. The female-only, exploratory male-only, and sex-matched analyses were performed as complementary analyses to assess the robustness of the findings with respect to the observed sex imbalance rather than to replace the primary validation strategy.

2.2. Participants and Measures

The study included 88 adult psychiatric inpatients recruited from two psychiatric hospitals in Romania between December 2023 and September 2024. Participation was voluntary, and all participants provided written informed consent prior to enrollment. Participants were divided into two groups based on clinical status at the time of recording. Group 1 comprised 59 patients admitted following a recent suicide attempt. All patients fulfilled the diagnostic criteria for severe depressive disorder at the time of hospitalization. Group 2 comprised 29 patients with severe depressive disorder and no clinically documented suicidal ideation or recent suicidal behavior at the time of inclusion. The classification into the two clinical groups was established by the treating psychiatrists based on the complete clinical assessment performed at hospital admission, including psychiatric examination, clinical history, medical records, and routine suicide risk evaluation. We acknowledge that the assessment of suicidal ideation inevitably relies partly on patient self-report and therefore cannot be considered completely independent from the limitation that motivated the present study. All participants were adults aged 18–65 with a primary diagnosis of severe depression (Table 1).

Table 1. Demographic characteristics of the study sample.

Characteristic	Suicide Attempters (<i>n</i> = 59)	Depression (<i>n</i> = 29)	Total (<i>N</i> = 88)
Male, <i>n</i> (%)	29 (49.2%)	5 (17.2%)	34 (38.6%)
Female, <i>n</i> (%)	30 (50.8%)	24 (82.8%)	54 (61.4%)
Age range (years)	18–65	18–65	18–65

The two clinical groups differed significantly in sex distribution (49.2% male in the suicide-attempt group versus 17.2% male in the depression group; χ^2 test, $p < 0.01$). Because sex influences several acoustic characteristics of speech, sex imbalance represented a potential source of confounding. Accordingly, in addition to the primary repeated patient-level cross-validation analysis, female-only, exploratory male-only, and repeated sex-matched analyses were performed to evaluate the robustness of the observed classification performance after partially controlling for sex. The results of these analyses are presented in Section 4.4. Participants were informed about the study’s objectives, stages, and data protection measures, and were assured of their right to withdraw at any time. All subjects agreed to be included in the study and signed the informed consent form before participating. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committees of both hospitals (Approval no. 39420/07.12.2023 from the Ethics Committee of “Prof. Dr. Al. Obregia” Psychiatry Clinical Hospital, Bucharest, Romania, and Approval no. 2/29.01.2024 from the Ethics Committee of “Elisabeta Doamna” Psychiatry Hospital, Galati, Romania). Data was collected from December 2023 to September 2024 and analyzed from November 2024 onward.

Audio recordings were obtained using Sony ICD-PX240 digital voice recorders, bought in Bucharest and Galati, at both hospital sites. Recordings were acquired in quiet clinical consultation rooms at a sampling rate of 44.1 kHz and encoded in MP3 format at 128 kbps, reflecting the routine recording protocol used in the participating hospitals. Each recording lasted between 8 and 12 min per patient. The same recording equipment and acquisition protocol were used at both sites to minimize recording-condition variability. Although MP3 compression represents a limitation for detailed acoustic analysis, its influence was minimized using an identical acquisition protocol for all participants. Future prospective studies will employ lossless audio acquisition to preserve the full acoustic information.

The interview protocol was semi-structured and consisted of two parts. The first part was an assessment of the suicide attempt (for Group 1) or of the patient’s current emotional state (for Group 2). The second part consisted of a discussion about emotionally neutral topics such as weather and hobbies, with the aim of obtaining speech under both emotionally salient and emotionally neutral conditions. Although interviewer speech was completely removed before acoustic analysis through the speaker-diarization procedure described below, interviewer-related effects cannot be completely excluded because interviewers were aware of participants’ clinical status and the interview topics differed between the two clinical groups. These differences may have influenced patients’ emotional arousal, conversational dynamics, or topic-related prosody. Consequently, the present findings should be interpreted as demonstrating acoustic separability between patient speech recorded under these interview conditions rather than establishing the existence of a unique suicide-specific vocal biomarker. Future multicenter studies should employ standardized interview protocols with identical elicitation tasks across groups together with blinded or automated interview administration. Interviewers were not blinded to patient group status; however, the technical team responsible for audio processing, model development, training, and

evaluation received only anonymized recordings and had no access to patient identifiers or clinical metadata other than the binary class labels required for supervised learning.

Before acoustic preprocessing, all interview recordings underwent automatic speaker diarization using the pretrained pyannote/speaker-diarization-3.1 model. Separate audio files were automatically generated for each detected speaker. Every diarized recording was subsequently reviewed manually by the research team to verify correct speaker assignment. This manual verification was performed before any acoustic preprocessing to ensure that only patient speech entered the machine-learning pipeline. Only the speech segments corresponding to the patient were retained for all subsequent analyses, whereas interviewer speech was completely excluded before silence removal, spectrogram generation, CNN training, and model evaluation. Consequently, all mel-spectrogram generation, CNN training, validation, and performance evaluation reported in this study were performed exclusively on patient speech.

Medication data for the participants was not systematically recorded and is therefore unavailable. This is acknowledged as a limitation, as psychotropic medications, including antipsychotics, benzodiazepines, and mood stabilizers, may influence voice production through effects on psychomotor activity, respiratory control, phonation, and laryngeal muscle tone, and therefore represents an important source of biological variability.

To prevent speaker-level data leakage, all speech segments extracted from a given participant were always assigned to the same cross-validation fold. Patient-level stratified cross-validation was performed before audio segmentation, ensuring that all recordings from an individual participant remained exclusively within either the training or the testing partition during each fold. Consequently, no speech segments originating from the same participant were present simultaneously in both the training and testing datasets. This strategy ensured complete independence between the training and testing data throughout all repeated cross-validation analyses.

2.3. Feature Representation and Classification Model

2.3.1. The Neurophysiological Basis of Emotional Signatures in Voice

Voice production is governed by a complex interplay between the central and autonomic nervous systems, involving the coordinated activity of respiratory muscles, the larynx, and the supralaryngeal vocal tract. Research on voice stress analysis has established that the autonomic nervous system shapes multiple acoustic parameters, including fundamental frequency, voice quality, and spectral energy distribution, through mechanisms that are partially independent of the speaker's conscious control [22]. The larynx is innervated by branches of the vagus nerve (cranial nerve X), which serves as the primary parasympathetic pathway between the brainstem and the visceral organs, while sympathetic innervation reaches the laryngeal musculature via the superior cervical ganglion and the external branch of the superior laryngeal nerve. During states of emotional arousal, autonomic activation modulates subglottal air pressure, vocal fold tension, and laryngeal positioning, all of which directly influence measurable acoustic parameters including fundamental frequency (F0), jitter, shimmer, harmonic-to-noise ratio, and spectral energy distribution [22].

These autonomic influences on voice production operate substantially below the threshold of volitional control. The anterior cingulate cortex (ACC) plays a central role in both emotional processing and vocal motor regulation, mediating the interaction between bottom-up arousal signals and top-down regulatory processes. Van Puyvelde et al. proposed the Model for Voice and Effort (MoVE), which integrates these two streams: F0 values and jitter reflect bottom-up arousal activity, while their range and variance are modulated by ongoing top-down control [22]. The periaqueductal grey matter (PAG), a key subcortical

structure in this pathway, interfaces with the autonomic pontomedullary structures—the parabrachial complex, the Kolliker–Fuse nucleus, and the nucleus tractus solitarius—that shape the cardiorespiratory motor patterns essential for vocalization [22]. Crucially for clinical applications, stress-induced changes in intrinsic laryngeal muscle activity have been shown to occur even in vocally healthy individuals exposed to psychological stressors, with both personality traits and vagal tone predicting the magnitude of these involuntary vocal changes [53].

This neurophysiological architecture has direct implications for psychiatric assessment. When a patient denies suicidal ideation during a clinical interview, the semantic content of speech is largely under conscious control. However, several acoustic characteristics of voice production continue to reflect autonomic and neurophysiological processes that are only partially subject to voluntary regulation. Consequently, voice analysis may provide complementary information about the speaker’s emotional state that is not fully captured by the semantic content of speech alone. The emotional state of the speaker is encoded in the spectral and temporal microstructure of the voice signal through mechanisms that are, by their neural architecture, partially involuntary. This principle forms the scientific basis for using voice analysis as a non-self-report biomarker: it accesses a channel of emotional expression that operates in parallel with, and partially independent of, the semantic content of speech. The present study does not attempt to infer suicidal intent directly from speech. Instead, it investigates whether deep learning can detect reproducible acoustic patterns associated with patients who recently attempted suicide when compared with a clinically relevant control group consisting of patients with severe depression without suicidal ideation.

2.3.2. Emotional States in the Pre-Suicidal Period

The emotional phenomenology of the pre-suicidal state is characterized by intense ambivalence, psychological pain, and dysregulated arousal. Clinical interviews with suicide attempt survivors in the present study consistently revealed descriptions of overwhelming inner turmoil preceding the act, a coexistence of the wish to die and the wish for the unbearable pain to cease, which is not identical to a straightforward desire for death. This ambivalence generates a distinctive emotional signature that is qualitatively different from the affective flattening and psychomotor retardation typical of depression without active suicidal behavior.

Neurobiological evidence supports the existence of distinct physiological states in individuals who have attempted suicide compared with those experiencing depression alone. DeVille et al. demonstrated that suicide attempters exhibit an “interoceptive numbing” pattern, characterized by increased tolerance for aversive sensations (longer breath-hold and cold-pressor tolerance) and decreased accuracy in perceiving non-aversive bodily signals (lower heartbeat perception accuracy), mirrored by reduced activation of the dorsal mid/posterior insular cortex [54]. This pattern, which was specifically linked to the act of attempting suicide, independent of psychiatric diagnosis or medication exposure, suggests that suicidal behavior is accompanied by measurable physiological changes that extend beyond the mood disturbances associated with depression. The coexistence of intense emotional distress with blunted interoceptive awareness creates a paradoxical internal state that, we hypothesize, produces vocal signatures distinguishable from those of depressive affect alone.

Furthermore, the temporal dynamics of emotional arousal in the context of suicidal crises suggest that these signatures persist in the voice beyond the immediate moment of crisis. Evidence from Romanian speech datasets indicates that negative arousal can persist in vocal parameters for minutes after an emotionally triggering interaction and

may intensify as a stressful event is anticipated [42]. The available evidence supports persistence on the order of minutes [42], while the recordings in the present study were obtained hours to days after the suicide attempt. The extent to which the observed acoustic differences reflect the persistence of acute emotional arousal versus a more stable post-crisis physiological state remains an open question. In the clinical context, patients interviewed within hours or days of a suicide attempt are likely to carry residual emotional activation that may be detectable through analysis of vocal acoustics, even when their verbal reports suggest resolution or improvement.

Importantly, the present study does not attempt to demonstrate that any individual acoustic feature is specific to suicidal behavior. Rather, it investigates whether the combination of multiple spectro-temporal characteristics, learned automatically by a deep neural network, contains sufficient information to distinguish patients with recent suicide attempts from clinically comparable patients with depression without suicidal ideation.

2.3.3. Spectrogram Representation

The acoustic signal of human speech is a complex waveform whose emotion-related information is distributed across both temporal and spectral dimensions. To make this information suitable for deep learning, the audio recordings were transformed into mel-spectrograms, two-dimensional time-frequency representations that convert the speech analysis problem into an image classification task while preserving the spectral characteristics relevant to emotional expression.

In the present study, the original recordings (44.1 kHz, MP3 format) were first resampled to 22,050 Hz. As described in Section 2.2, all recordings had previously undergone automatic speaker diarization followed by manual verification, and only patient speech was retained for all subsequent analyses. Silent intervals and non-speech regions were then removed using an energy-based voice activity procedure (`librosa.effects.split, top_db = 20`) to reduce the influence of pauses and background silence. The remaining speech signal was divided into two-second segments with 50% overlap between consecutive segments. For each segment, a short-time Fourier transform (STFT) was computed using an FFT window size of `n_fft = 2048` samples (approximately 93 ms at 22,050 Hz) and a hop length of 512 samples (approximately 23 ms). The resulting power spectra were projected onto 128 mel-frequency bands covering the frequency range from 0 to 11,025 Hz, thereby generating a mel-spectrogram for each speech segment.

Each mel-spectrogram was converted to decibel units using `librosa.power_to_db` and subsequently standardized using Z-score normalization (mean subtraction followed by division by the standard deviation). Spectrograms shorter than 100 time frames were zero-padded, whereas longer spectrograms were truncated to 100 frames, resulting in a fixed input size of $128 \times 100 \times 1$ (128 mel-frequency bands $\times$ 100 time frames $\times$ one grayscale channel) for all CNN inputs.

Because silent intervals were removed before segmentation, the number of speech segments generated from each participant varied according to the duration of voiced speech within the interview. Consequently, individual participants contributed different numbers of spectrograms during model training. Importantly, all segments originating from a given participant were always assigned to the same data partition, ensuring complete patient-level separation between the training, validation, and testing datasets and preventing speaker-level information leakage.

Although the CNN was trained using individual spectrogram segments, all performance metrics reported in this study were computed at the patient level. Segment-level predictions belonging to the same participant were aggregated using majority voting to determine the final patient classification. Because receiver operating characteristic (ROC)

analysis requires a continuous prediction score, the mean predicted probability of the suicide-attempter class across all segments of a participant was used to calculate the patient-level ROC-AUC. The repeated patient-level stratified five-fold cross-validation described in Section 2.2 was applied to the entire preprocessing and classification pipeline. For every fold, all preprocessing steps, including speech segmentation, mel-spectrogram generation, model training, and patient-level evaluation, were performed independently using only the participants assigned to the corresponding training and testing partitions. This procedure ensured complete separation of participants between folds and prevented any form of speaker-level information leakage throughout the validation process.

The mel scale provides a perceptually motivated frequency representation that approximates the nonlinear frequency resolution of the human auditory system by expanding lower frequencies, where auditory discrimination is greatest, and compressing higher frequencies. This representation has consistently demonstrated superior performance compared with linear-frequency spectrograms in speech emotion recognition because it preserves the spectral regions most strongly associated with prosody, phonation, and vocal tract resonances [38,55].

Mel-spectrograms are particularly well suited for convolutional neural networks because they transform complex acoustic phenomena into spatial patterns that can be learned hierarchically. Emotion-related variations in pitch, phonation stability, formant trajectories, and spectral energy distribution appear as characteristic time-frequency patterns that CNNs can model through successive convolutional layers. Rather than relying on predefined acoustic features, the proposed Multi-Scale CNN automatically learns multiscale spectrotemporal representations that capture complex combinations of acoustic characteristics associated with the studied clinical groups. Representative mel-spectrograms from both clinical groups are shown in Figure 1.

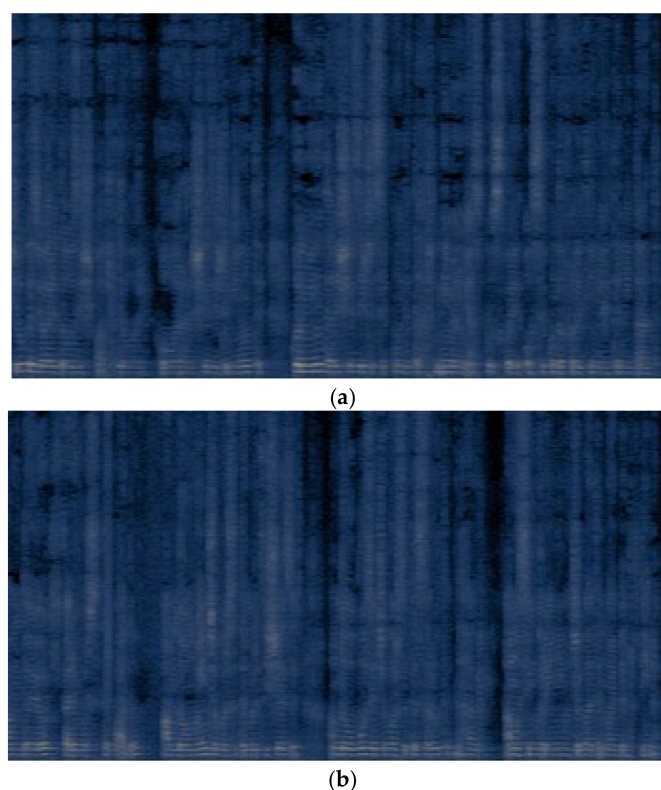


Figure 1. (a) Representative mel-spectrogram example of a patient with a survived suicide attempt, interviewed on the first day of hospitalization. (b) Representative mel-spectrogram example of a patient with depression, interviewed on the first day of hospitalization.

2.3.4. Model Architecture: Multi-Scale Convolutional Neural Network with Attention

The classification model employed in this study is a multi-scale two-dimensional convolutional neural network (Multi-Scale CNN) with residual connections and spatial attention, implemented using the Keras Functional API. The model processes mel-spectrograms as single-channel grayscale images ($128 \times 100 \times 1$). The architecture comprises four principal stages: multi-scale feature extraction, convolutional processing with a residual connection, a spatial attention mechanism, and a classification head. The overall architecture is illustrated in Figure 2. The same CNN architecture and hyperparameters were used throughout all experiments reported in this study, including the overall repeated patient-level stratified five-fold cross-validation, the female-only analysis, the exploratory male-only analysis, and the sex-matched analysis. Only the composition of the patient cohorts differed between these experiments, allowing the robustness of the learned acoustic representations to be evaluated under different levels of control for sex imbalance.

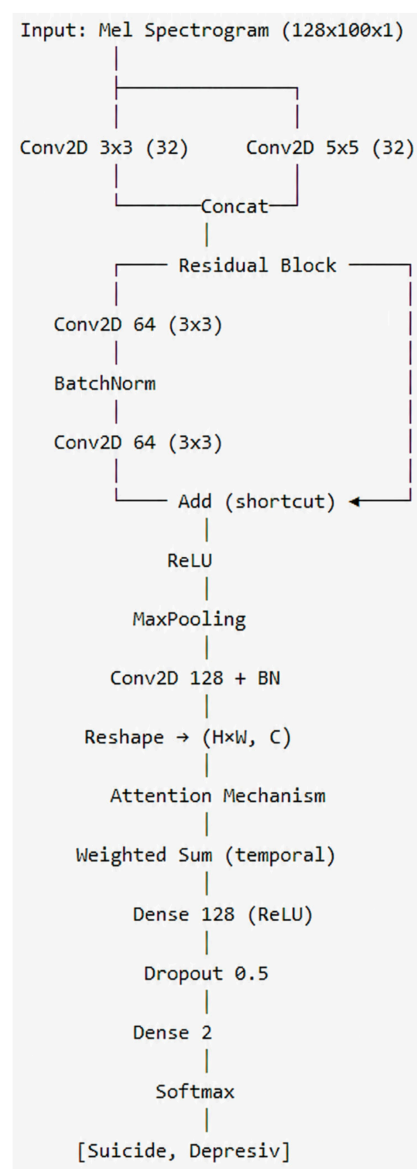


Figure 2. CNN architecture.

(Schematic diagram of the Multi-Scale CNN architecture with residual connection and spatial attention mechanism. The network processes mel-spectrograms ($128 \times 100 \times 1$) through parallel 3×3 and 5×5 convolutional branches, concatenates their outputs, and

feeds them into a residual block (a 3×3 convolution with a 1×1 projection shortcut via skip connection) followed by ReLU activation and max pooling. A subsequent 128-filter convolution with batch normalization is processed through a spatial attention mechanism (reshape $\rightarrow$ attention weighting $\rightarrow$ temporal weighted sum), before passing through a fully connected layer (128 units, ReLU), dropout (0.5), and the binary softmax output [suicide, depression]).

Multi-scale feature extraction. Two parallel convolutional branches process the input simultaneously. The first branch applies 32 filters with a (3×3) kernel to capture fine-grained spectro-temporal patterns, while the second branch applies 32 filters with a (5×5) kernel to capture broader context. The outputs of both branches are concatenated along the channel dimension, yielding a 64-channel feature map ($128 \times 100 \times 64$). This multi-scale design allows the network to simultaneously learn local acoustic details and wider contextual patterns from the mel-spectrogram.

Convolutional processing with residual connection. The concatenated features are processed by a Conv2D layer (64 filters, 3×3 kernel) followed by batch normalization. A residual block then applies two parallel paths: a Conv2D layer (64 filters, 3×3 kernel) for nonlinear transformation and a Conv2D layer (64 filters, 1×1 kernel) serving as a linear projection shortcut. The outputs are summed (residual connection) and passed through ReLU activation and max pooling (2×2). This residual design mitigates the vanishing gradient problem and facilitates training of deeper representations. A subsequent Conv2D layer (128 filters, 3×3 kernel) with batch normalization further refines the feature maps.

Spatial attention mechanism. The feature maps are reshaped and processed through a dense layer (1 unit) followed by softmax activation to compute attention weights over the spatial dimensions. These weights are multiplied element-wise with the feature representation, producing an attention-weighted summary vector (128 dimensions). This mechanism allows the model to focus on the most discriminative spectro-temporal regions of the mel-spectrogram.

Classification head. The attention output is passed through a fully connected layer (128 units, ReLU) with dropout (rate 0.5), followed by the output layer (2 units, softmax), producing probability estimates for each class.

The parameter counts per layer are as follows: parallel Conv2D branches, $320 + 832 = 1152$; Conv2D (64, 3×3) + BatchNorm, $36,928 + 256 = 37,184$; residual block (Conv2D $64 \times 3 \times 3$ + Conv2D $64 \times 1 \times 1$), $36,928 + 4160 = 41,088$; Conv2D (128, 3×3) + BatchNorm, $73,856 + 512 = 74,368$; attention (Dense 1), 129; Dense (128), 16,512; output Dense (2), 258. The total parameter count is 170,691 (170,307 trainable parameters and 384 non-trainable parameters from batch normalization running statistics).

The Multi-Scale CNN approach treats mel-spectrograms as images, exploiting the spatial structure inherent in time-frequency representations. Horizontal patterns in the spectrogram correspond to temporal dynamics, capturing how spectral content changes over time. Vertical patterns capture harmonic structure, representing frequency relationships at a given instant. Diagonal patterns can capture pitch contours and formant transitions. This image-based paradigm for processing spectrograms is well-established in speech emotion recognition [38,55].

The multi-scale feature extraction enables hierarchical feature learning: the (3×3) branch detects fine-grained spectro-temporal edges, while the (5×5) branch captures broader patterns spanning wider frequency and time ranges. The residual connection facilitates gradient flow during training and enables the network to learn identity mappings where appropriate, reducing the risk of degradation with depth. The spatial attention mechanism provides interpretability by learning to weight the most discriminative regions of the spectrogram, allowing the model to focus on spectro-temporal areas most relevant to classi-

fication. BatchNormalization stabilizes training by normalizing intermediate activations and acts as a mild regularizer. Dropout (0.5) provides regularization in the classification head, where overfitting risk is greatest. No architectural modifications or hyperparameter tuning were performed separately for the additional sex-controlled analyses, ensuring that all comparisons reflected differences in patient composition rather than differences in model design.

The model was implemented in TensorFlow/Keras (version 2.x) and trained using the Adam optimizer with an initial learning rate of 2×10^{-4} . ReduceLROnPlateau was applied, reducing the learning rate when validation loss plateaued (observed reductions to 2×10^{-5} and 2×10^{-6} during training). Training was conducted with a batch size of 16 for 20 epochs. No early stopping was used, and no data augmentation techniques (such as noise injection, pitch shifting, or time stretching) were applied. The data were split into training and test sets using a stratified train/test split (train_test_split with stratify = y) to preserve class proportions in both sets. No explicit class weighting was employed despite the 2:1 class imbalance (59 suicide attempters vs. 29 depression patients).

2.3.5. Binary Classification Rationale

The present study employs a binary classification scheme, distinguishing between patients who recently attempted suicide (Group 1, $n = 59$) and patients with severe depression without suicidal ideation (Group 2, $n = 29$). This binary design was chosen to establish, as a first step, whether the Multi-Scale CNN model can reliably separate the acoustic patterns of recent suicide attempters from those of depressed patients without suicidal ideation. Although the original cohort presented an approximately 2:1 class imbalance together with a significant sex imbalance, the revised study additionally includes female-only, exploratory male-only, and sex-matched analyses to evaluate the robustness of the proposed approach under different cohort compositions.

This distinction is clinically meaningful because it tests whether suicidal behavior leaves a distinct vocal signature beyond what depression itself produces. Depression is associated with well-documented changes in vocal characteristics, including reduced pitch variability, slower speech rate, and altered spectral energy distribution. By comparing suicide attempters against patients with severe depression who do not report suicidal ideation, the design controls for the acoustic effects of depression per se and asks whether the act of attempting suicide, and the physiological and emotional states surrounding it, introduces additional acoustic markers that a model can detect. If the model can classify patients on this dimension from voice alone, it would suggest that suicidal behavior is accompanied by vocal changes that go beyond those attributable to depressive illness, representing a step toward identifying patients whose acoustic profiles indicate acute suicide-related distress rather than depression alone.

Extension to a multi-level classification system, incorporating gradations of risk intensity such as residual vulnerability in treated patients, is planned as a future direction once larger samples are available.

2.4. Declaration of Generative AI Use

During the preparation of this manuscript, the authors used Claude (Anthropic; Claude Opus 4.7) to assist with the following tasks: (i) drafting, rewriting, and refining the manuscript text, including the structured abstract, figure captions, Acknowledgments, and Data Availability Statement; (ii) literature search support, including querying PubMed and the Consensus academic database to identify relevant publications; (iii) verification of citation metadata (author lists, DOIs, volumes, and page numbers) against primary sources; (iv) computation and interpretation of confidence intervals, exact binomial test

p-values, and the sex-only baseline AUC reported in Section 4.4; and (v) formatting of references into the MDPI style. The tool was not used to design the study, to recruit or interview participants, to collect or analyze the primary audio data, to develop or train the convolutional neural network model, to generate any of the figures, or to fabricate any reference or result. All analytical decisions, model development, empirical results, and final scientific content are the work of the human authors. After using this tool, the authors reviewed, verified, and edited all content as needed and take full responsibility for the content of the publication.

3. Results

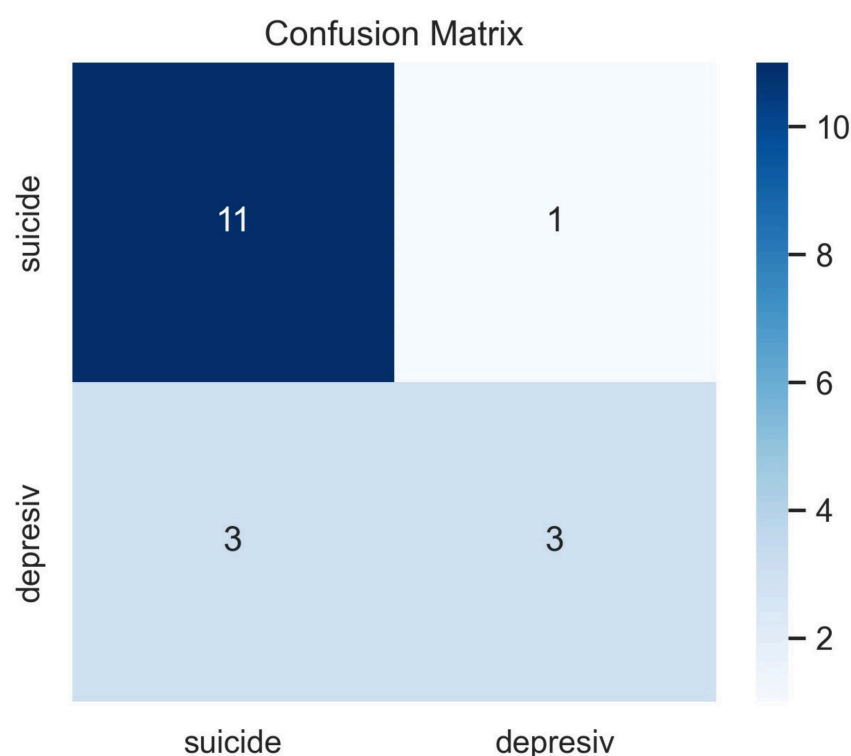
The revised evaluation was performed using repeated patient-level stratified five-fold cross-validation on the complete dataset of 88 psychiatric inpatients (59 suicide attempters and 29 patients with severe depression without suicidal ideation). Following speaker diarization and preprocessing, mel-spectrograms were generated from two-second speech segments with 50% overlap. All speech segments originating from the same participant were assigned exclusively to a single fold, thereby ensuring complete patient-level separation between training and testing data and preventing speaker-level information leakage. For each fold, the entire preprocessing pipeline, spectrogram generation, CNN training, and patient-level evaluation were performed independently. Final patient-level classification was determined by majority voting across all speech segments belonging to the same participant, whereas patient-level ROC-AUC values were computed using the mean predicted probability of the suicide-attempter class across all segments of the corresponding participant. Across the repeated patient-level stratified five-fold cross-validation, the proposed Multi-Scale CNN achieved a mean patient-level classification accuracy of $72.67\% \pm 6.44\%$, a balanced accuracy of $61.83\% \pm 9.56\%$, and a mean ROC-AUC of 0.824 ± 0.054 . Mean sensitivity for suicide attempters was 0.883 ± 0.095 , whereas mean specificity for patients with depression was 0.413 ± 0.250 . The higher sensitivity indicates that the model consistently detected the clinically important suicide-attempter group across the different validation folds, although at the expense of a lower specificity for the depression group. Per-class performance metrics are summarized in Table 2. To investigate the potential influence of sex imbalance, additional analyses were performed using female-only, exploratory male-only, and sex-matched cohorts. The female-only analysis produced performance comparable to that observed in the complete cohort (mean ROC-AUC = 0.866 ± 0.123), whereas the exploratory male-only analysis yielded a mean ROC-AUC of 0.786 but should be interpreted cautiously because of the limited number of male participants. In the sex-matched analysis, the mean ROC-AUC decreased to 0.621, indicating that part of the discriminative performance may be influenced by sex-related acoustic differences. Nevertheless, classification performance remained above chance, supporting the existence of additional acoustic information beyond sex alone. The results demonstrate that the proposed Multi-Scale CNN consistently identified suicide attempters with higher sensitivity than patients with depression across the repeated validation folds. This operating point favors the detection of clinically relevant high-risk individuals but also produces a relatively high false-positive rate among depressed patients. Consequently, the proposed approach should be regarded as a complementary decision-support tool intended to assist, rather than replace, comprehensive clinical assessment.

Table 2. Mean patient-level classification performance obtained by repeated stratified five-fold cross-validation (mean $\pm$ standard deviation).

Metric	Mean $\pm$ SD
Accuracy	72.67 $\pm$ 6.44
Balanced accuracy	61.83 $\pm$ 9.56
ROC-AUC	0.824 $\pm$ 0.054
Sensitivity	0.883 $\pm$ 0.095
Specificity	0.413 $\pm$ 0.250
Precision (suicide)	0.761 $\pm$ 0.070
Macro F1	0.633 $\pm$ 0.142

Values are reported as the mean $\pm$ standard deviation across the five patient-level cross-validation folds.

Across the repeated patient-level five-fold cross-validation, the proposed Multi-Scale CNN consistently achieved higher sensitivity than specificity. Mean sensitivity for the suicide-attempter class reached 0.883 ± 0.095 , whereas mean specificity for the depression group was 0.413 ± 0.250 . This operating point favors the identification of clinically relevant high-risk individuals but also results in a relatively high false-positive rate among patients with depression. Consequently, the proposed approach should be regarded as a complementary decision-support tool intended to assist, rather than replace, comprehensive clinical assessment. A representative confusion matrix obtained from one cross-validation fold is presented in Figure 3.

**Figure 3.** Representative patient-level confusion matrix obtained from one fold of the repeated patient-level stratified five-fold cross-validation. Final patient classifications were determined by majority voting across all speech segments belonging to each participant.

Performance variability was assessed through repeated patient-level stratified five-fold cross-validation. The standard deviations reported in Table 2 quantify the variability

associated with different training and testing partitions. Although this revised validation strategy provides substantially more robust performance estimates than a single train/test split, the observed variability indicates that larger multicenter datasets remain necessary for precise estimation of clinical performance.

For each cross-validation fold, the Multi-Scale CNN was trained independently using the same architecture and hyperparameters. The training curves demonstrated stable convergence across folds. Training accuracy approached approximately 97–98%, whereas validation performance stabilized at substantially lower values, indicating persistent but moderate overfitting despite the regularization techniques incorporated into the network architecture. Representative training and validation accuracy and loss curves from one fold are shown in Figure 4.

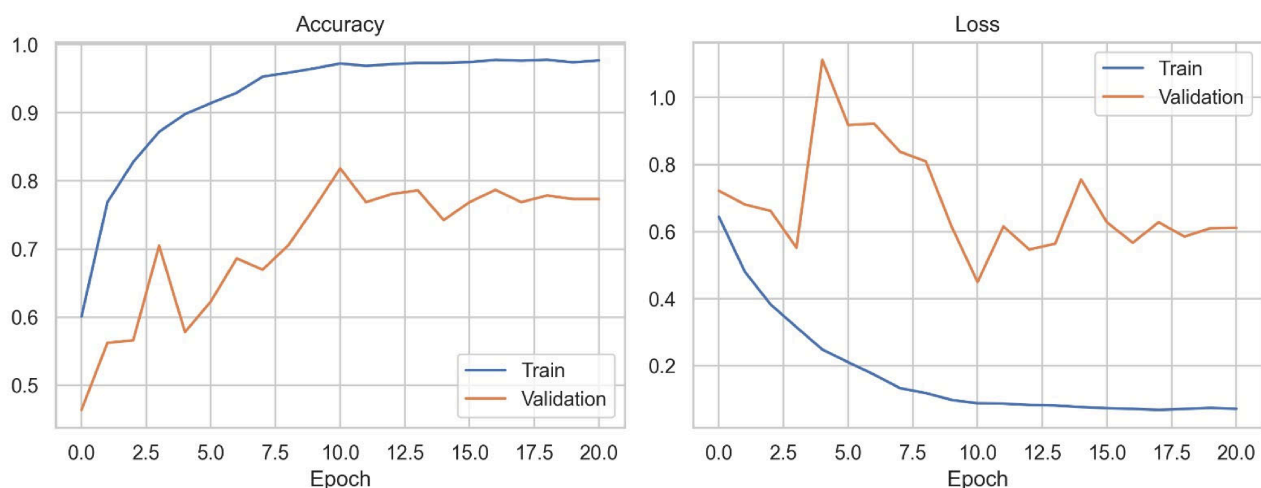


Figure 4. Representative training and validation accuracy and loss curves obtained from one fold of the repeated patient-level stratified five-fold cross-validation. The curves illustrate the typical convergence behavior of the proposed Multi-Scale CNN.

The representative spectrograms shown in Figure 1 are intended solely to illustrate the preprocessing pipeline and the mel-spectrogram representation used as input to the CNN. They are not expected to exhibit visually obvious class-specific differences. Instead, discrimination emerges from the automatic learning of complex multiscale spectro-temporal patterns across thousands of spectrogram segments.

As a conventional machine-learning baseline, a Random Forest classifier trained on Mel-frequency cepstral coefficient (MFCC) features was evaluated using exactly the same patient-level five-fold cross-validation partitions as the proposed Multi-Scale CNN. The Random Forest achieved a mean accuracy of $75.0 \pm 3.4\%$, a mean ROC-AUC of 0.818 ± 0.123 , a sensitivity of 0.950 ± 0.075 , a specificity of 0.340 ± 0.107 , and a macro-F1 score of 0.651 ± 0.056 . In the pooled patient-level comparison, the Random Forest and the Multi-Scale CNN achieved accuracies of 75.0% and 72.7%, respectively. An exact McNemar test showed no statistically significant difference between the paired patient-level classifications of the two models ($p = 0.804$), indicating that both approaches achieved comparable overall performance on the present dataset.

The detailed comparison between the proposed Multi-Scale CNN and the conventional MFCC-based Random Forest baseline is summarized in Table 3.

Table 3. Comparison between the proposed Multi-Scale CNN and the MFCC-based Random Forest baseline.

Metric	Multi-Scale CNN	MFCC + Random Forest
Accuracy	72.7 ± 6.4	75.0 ± 3.4
ROC-AUC	0.824 ± 0.054	0.818 ± 0.123
Sensitivity	0.883 ± 0.095	0.950 ± 0.075
Specificity	0.413 ± 0.250	0.340 ± 0.107
Macro-F1	0.633 ± 0.142	0.651 ± 0.056
McNemar <i>p</i> -value	colspan	0.804
Balanced accuracy	61.8 ± 9.6	64.5 ± 4.8

The parameter-to-patient ratio for the Multi-Scale CNN is $170,691/88 =$ approximately 1940:1, which is discussed further in Section 4.6.

Across the repeated patient-level stratified five-fold cross-validation, the proposed Multi-Scale CNN achieved a mean ROC-AUC of 0.824 ± 0.054 (Table 2), indicating good discriminative ability across different patient partitions. A representative ROC curve obtained from one cross-validation fold is shown in Figure 5, whereas the precision–recall curve for the same representative fold is presented in Figure 6.

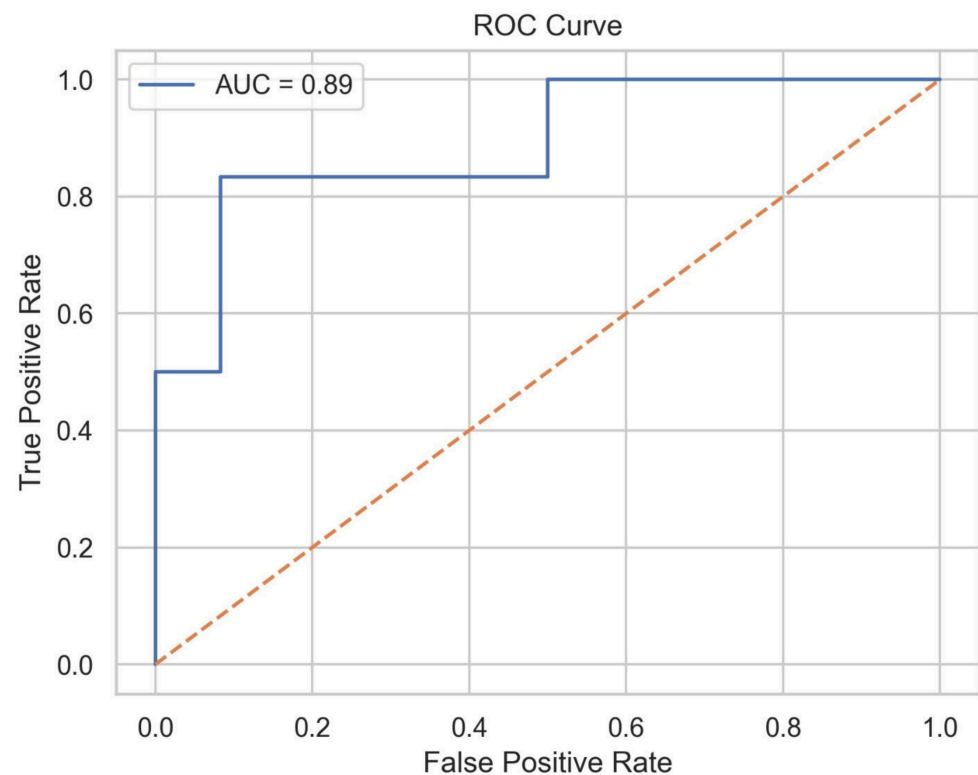


Figure 5. Representative receiver operating characteristic (ROC) curve obtained from one fold of the repeated patient-level stratified five-fold cross-validation. The ROC-AUC shown corresponds to a representative fold and is presented for illustration. Overall patient-level performance is summarized in Table 2 using the mean ROC-AUC across all cross-validation folds.

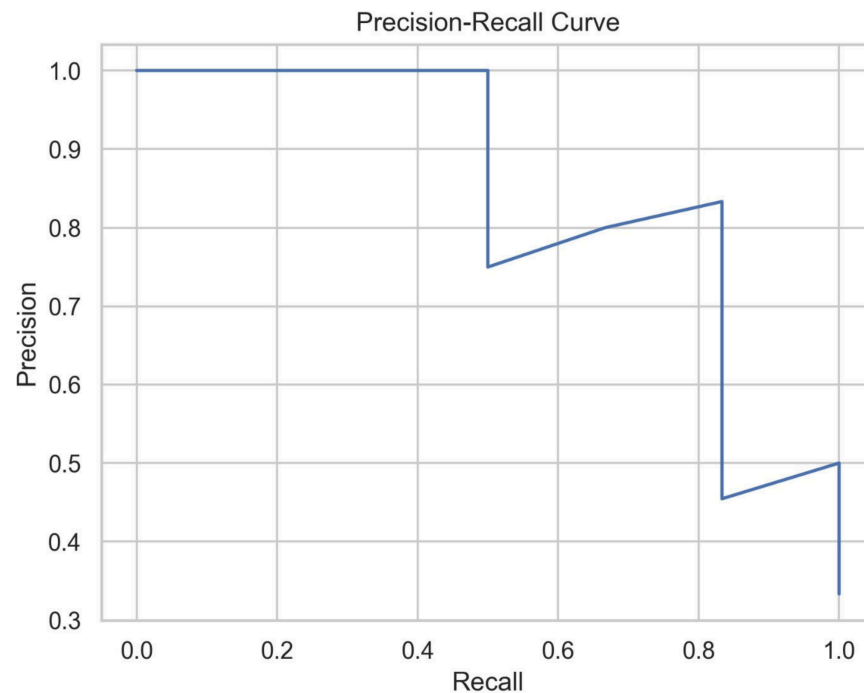


Figure 6. Representative precision–recall curve obtained from one fold of the repeated patient-level stratified five-fold cross-validation.

Additional performance measures, including Cohen’s kappa, calibration analysis, and permutation-based significance testing, were not included in the present proof-of-concept study and will be considered in future validation studies involving larger multicenter datasets.

4. Discussion

4.1. The Dissimulation Problem and the Case for Non-Self-Report Biomarkers

The results of this study should be interpreted against the backdrop of a fundamental clinical challenge: patients at the highest risk of suicide frequently deny their suicidal intent. As reported in Section 1, the empirical literature consistently documents denial rates of approximately 50% among individuals with active suicidal ideation [14], with two-thirds of patients who died by suicide while in care having denied ideation at their last clinical contact [15], and 78% of inpatient suicide decedents denying ideation at their last communication before death [16].

These findings expose a structural vulnerability in all assessment instruments that depend on patient self-report, including the Columbia-Suicide Severity Rating Scale (C-SSRS), the Beck Scale for Suicide Ideation, the Ask Suicide-Screening Questions (ASQ), and the Patient Health Questionnaire Item 9. While these tools have demonstrated strong psychometric properties when patients respond truthfully—the ASQ, for example, achieves sensitivity of 97% and high negative predictive values [56]—their clinical utility is fundamentally constrained by the gate of patient disclosure. When the most acutely at-risk patients are also the most likely to deny their ideation, the predictive architecture of self-report instruments is undermined precisely where it is needed most.

Voice-based biomarker analysis addresses this vulnerability through a fundamentally different measurement channel. As described in the preceding section, the acoustic properties of speech are shaped by autonomic nervous system activity that operates partially below volitional control. The laryngeal muscles respond to sympathetic and parasympathetic activation patterns driven by subcortical structures—the periaqueductal grey, the

parabrachial complex, the nucleus tractus solitarius—that mediate the physiological stress response independently of the speaker’s conscious communicative intent [22]. A patient may choose to say, “I am not thinking about suicide,” but the spectral energy distribution, jitter, pitch variability, and formant characteristics of the voice carrying that statement reflect an emotional state that the patient cannot fully suppress.

It is important to note that the present study design does not directly test the detection of actively concealed suicidal intent. Both groups were aware of their clinical status during interviews, and Group 1 patients were interviewed specifically about their suicide attempt; they were not denying or concealing their experience. The study demonstrates that the voices of recent suicide attempters are acoustically separable from those of depressed patients without suicidal ideation, which is a necessary precondition for dissimulation-resistant assessment. However, the sufficient condition—demonstrating that voice analysis can identify concealed suicidal intent in patients who are actively denying it—remains to be established in future prospective studies specifically designed for that purpose. The present repeated patient-level cross-validation analyses demonstrate that acoustic discrimination between recent suicide attempters and depressed controls is reproducible across multiple train-test partitions. Nevertheless, these findings should not be interpreted as evidence that concealed suicidal intent can already be detected in routine clinical practice.

The present study’s design, recording patients during clinical interviews that include both emotionally charged topics (the suicide attempt, emotional states prior to the act) and affectively neutral content (weather, hobbies), provides within-subject contrast that enhances the detection of emotion-specific vocal signatures. The inclusion of neutral speech topics allows the Multi-Scale CNN to learn the baseline vocal characteristics of each speaker, against which deviations during emotionally salient topics become more interpretable. This design feature is consistent with evidence that person-specific baselines matter when interpreting vocal change in the context of suicide risk [25]. Additional validation analyses were performed to evaluate the robustness of the proposed approach. Repeated patient-level stratified five-fold cross-validation yielded a mean classification accuracy of $72.7 \pm 6.4\%$ and a mean ROC-AUC of 0.824 ± 0.054 , providing a substantially more robust estimate of model performance than the original single train-test split. Female-only analyses maintained good discriminative performance (ROC-AUC 0.866 ± 0.123), indicating that meaningful discrimination persisted within a single sex. By contrast, repeated sex-matched analyses produced lower performance (ROC-AUC 0.621 ± 0.156), suggesting that sex imbalance contributed to, but did not fully explain, the observed discrimination. Because the male-only analysis included only five patients with depression, it was considered exploratory and should be interpreted cautiously. Overall, these additional analyses support a more balanced interpretation of the present findings. They indicate that the proposed Multi-Scale CNN captures reproducible acoustic differences between recent suicide attempters and patients with severe depression without documented suicidal ideation, while also demonstrating that part of the observed discrimination is influenced by sex-related variability. Consequently, the present findings should be interpreted as evidence supporting the feasibility of objective voice-based biomarkers rather than definitive identification of a suicide-specific vocal signature. Larger multicenter studies with balanced cohorts, standardized interview protocols, and external validation will be required before clinical implementation.

4.2. The Post-Discharge Period

The clinical context in which a voice-based monitoring tool would deliver the greatest benefit is the period following discharge from psychiatric hospitalization, a window that meta-analytic evidence consistently identifies as one of extraordinary and concentrated

suicide risk. It should be noted that this monitoring scenario represents a proposed clinical application that was not tested in the present study.

As reported in Section 1, post-discharge suicide rates rise to 1132 per 100,000 person-years within the first three months [15], with even higher rates in the earliest days: 6062 per 100,000 on days 0–3 [16] and 2950 per 100,000 in the first week [20]. Olfson et al. reported that among adults with depressive disorders in the US Medicaid population, the 90-day post-discharge suicide rate was 235 per 100,000 person-years [57]. Current clinical practice addresses this high-risk window primarily through scheduled follow-up appointments, crisis hotline availability, and in some settings, brief contact interventions. A meta-analysis of brief contact interventions in discharged psychiatric patients found significant reductions in suicide attempts within 12–14 months and in suicidal ideation within 12 months, though no significant effect on completed suicide was demonstrated [58]. However, these interventions are intermittent by nature and depend on the patient initiating contact or being reachable at predetermined intervals.

A voice-based monitoring system could offer a qualitatively different approach: daily, non-invasive assessment through interaction with a virtual assistant, where the patient's vocal patterns are analyzed in real time and changes in emotional intensity are flagged to the treating clinician. This model would leverage the temporal concentration of risk by providing the most intensive monitoring precisely when it is most needed, without requiring the patient to accurately self-report their emotional state. Existing feasibility data for psychiatric monitoring applications are encouraging: studies report retention rates of approximately 70% at four weeks for smartphone-based mental health monitoring applications, and risk ratios for vocal biomarker-based scores improve substantially with daily versus sporadic use [59]. The present findings should therefore be interpreted as supporting the feasibility of patient-level acoustic discrimination between recent suicide attempters and depressed patients without documented suicidal ideation, rather than demonstrating the effectiveness of a post-discharge monitoring system. Whether repeated voice analysis can improve early detection of suicidal crises during longitudinal follow-up remains an important question for future prospective studies.

4.3. Suicide Risk Assessment in Emergency Settings

The emergency department represents another clinical context where non-self-report biomarkers could provide significant value. It should be noted that this application scenario was not tested in the present study and represents a proposed direction for future research.

Population-based analyses show that 91.7% of suicide decedents had some form of healthcare contact in the year preceding their death, with the most common contact being an outpatient primary care visit (54.0%), followed by outpatient psychiatric visits (39.8%), emergency department visits (31.1%), and psychiatric hospitalizations (21.0%) [60]. The median time from last mental healthcare contact to death was 18 days [60]. These data indicate that clinical encounters occur with sufficient frequency to provide opportunities for risk detection, but the brevity and pressured nature of emergency department assessments, combined with the dissimulation problem, means that many at-risk patients pass through these encounters unidentified.

Universal suicide risk screening in emergency departments has shown promise for identifying patients who would otherwise be missed. Many emergency department patients who present with non-psychiatric complaints have elevated suicide risk, which universal screening can detect [61]. However, commonly used screening scales demonstrate only modest predictive performance for disposition decisions and subsequent suicidal behavior when used in isolation, reinforcing current recommendations that structured screening should be complemented by additional sources of objective clinical information. In a

prospective emergency department study, three widely used screening instruments (including the C-SSRS) displayed modest negative predictive values ranging from 0.66 to 0.73 [62]. The American College of Emergency Physicians clinical policy explicitly states that screening tools should not be used in isolation to guide disposition decisions.

Voice analysis could, in principle, augment emergency department screening by providing an objective, rapidly obtainable signal that does not depend on patient disclosure and does not add substantial time to the clinical encounter. A brief recording obtained during triage or routine clinical interaction, lasting no more than the time the patient already spends speaking with a clinician, could be processed in real time to generate a risk indicator that complements the structured screening instrument. This approach would be particularly valuable for the substantial proportion of patients who present with non-psychiatric complaints but carry unrecognized suicide risk, and for those who deny ideation when directly asked.

4.4. Alternative Explanations and Confounds

The observed classification performance should be interpreted considering several alternative explanations that could contribute to group separability without necessarily implying the existence of a suicide-specific vocal signature. The primary analysis based on repeated patient-level stratified five-fold cross-validation yielded a mean classification accuracy of $72.7 \pm 6.4\%$ and a mean ROC-AUC of 0.824 ± 0.054 , providing a more robust estimate of model performance than the single train-test split reported initially.

First, medication differences between the two groups represent a plausible confound. Patients admitted after a suicide attempt may have received emergency sedation, may have discontinued their medications prior to the attempt, or may have been in a different acute pharmacological state compared with patients hospitalized for severe depression without suicidal ideation. Since psychotropic medications, particularly antipsychotics and benzodiazepines, are known to affect psychomotor activity and vocal characteristics, systematic medication differences between groups could have contributed to the observed classification performance. Medication data were not systematically recorded in the present study, preventing assessment of the potential influence of psychotropic treatment on the observed acoustic differences. Future studies should systematically document medication exposure and incorporate it into multivariable analyses.

Second, differences in the time since admission and the time elapsed since the suicide attempt may contribute to group separability. Group 1 patients were recorded shortly after admission following an acute crisis, while Group 2 patients were hospitalized for depression and may have been recorded at various points during their stay. Acute post-crisis physiological arousal (elevated cortisol, catecholamines) and potential physical effects of the attempt itself (exhaustion, sedation, laryngeal effects from intubation or vomiting) could produce vocal differences unrelated to the emotional states of interest.

Third, interviewer-related effects remain a potential source of confounding, although interviewer speech itself was not used during model development. Before feature extraction, all recordings underwent automatic speaker diarization followed by manual verification, and only patient speech was retained for CNN training and evaluation. Consequently, the model could not learn interviewer-specific acoustic characteristics. However, interviewer-related effects cannot be completely excluded because interviewers were aware of participants' clinical status and the interview topics differed between the two clinical groups (recent suicide attempt versus current emotional state). These differences may have influenced patients' emotional arousal, conversational dynamics, or topic-related prosody. Therefore, the present findings should be interpreted as demonstrating acoustic separability between patient speech recorded under these interview conditions rather than

establishing the existence of a unique suicide-specific vocal biomarker. Future prospective studies should employ identical interview protocols across groups together with blinded or automated interview administration to minimize this potential source of confounding.

Fourth, depression severity may differ systematically between the two groups. Although both groups were diagnosed with severe depression, patients who have crossed the threshold to a suicide attempt may on average have more severe depression, greater hopelessness, and more profound psychomotor disturbance. If the Multi-Scale CNN is primarily detecting depression severity rather than a suicide-specific signal, the classification performance would represent severity detection with a correlated clinical label rather than evidence for a distinct vocal signature of suicidal behavior.

Fifth, the two groups differed substantially in sex distribution: Group 1 was approximately balanced (49.2% male) while Group 2 was predominantly female (82.8%). Sex differences in vocal characteristics are well-documented, with male fundamental frequency averaging approximately 120 Hz compared with approximately 170 Hz in females. These spectral differences are directly encoded in mel-spectrograms and are readily learnable by CNNs, which have been shown to classify speaker sex from spectrograms with accuracy exceeding 90% in published benchmarks. The observed classification performance may therefore partially reflect the model learning sex-related vocal characteristics that differ between groups rather than suicide-specific acoustic patterns. Following the reviewers' recommendations, additional analyses were performed to directly evaluate the influence of sex imbalance. Female-only analyses maintained good discriminative performance (mean ROC-AUC = 0.866 ± 0.123), indicating that meaningful acoustic discrimination persisted within a single sex. By contrast, repeated sex-matched analyses produced lower performance (mean ROC-AUC = 0.621 ± 0.156), demonstrating that sex imbalance contributed to, but did not fully explain, the observed discrimination. Because only five male patients were available in the depression group, the male-only analysis was considered exploratory and should be interpreted cautiously. These findings indicate that sex constitutes an important source of variability that should be carefully controlled in future multicenter studies, while arguing against the hypothesis that the observed discrimination is explained exclusively by sex differences.

These alternative explanations do not invalidate the finding that the two groups are acoustically separable, which is itself informative, but they indicate that the specific interpretation of what the model has learned remains uncertain. The additional repeated patient-level cross-validation analyses further indicate that the observed discrimination is reproducible across multiple train-test partitions rather than being attributable to a favorable random split, although the limited sample size still requires cautious interpretation. Disentangling these confounding factors will require larger prospective multicenter studies with balanced cohorts, standardized interview protocols, systematic recording of medication exposure, standardized timing of post-attempt recordings, external validation, and longitudinal follow-up. Such studies will be essential for determining which components of the learned acoustic representation reflect suicide-related neurophysiological alterations and which reflect correlated demographic or clinical variables.

4.5. Ethical Considerations

The development and potential deployment of voice-based biomarkers for suicide risk assessment raise several ethical issues that warrant explicit consideration. First, the capacity to provide informed consent may be compromised in patients experiencing acute psychiatric illness, including those in suicidal crisis. The present study obtained written informed consent from all participants, but any future deployment of a continuous monitoring system must establish clear protocols for consent that account for fluctuations in decisional

capacity, including provisions for advanced directives and surrogate decision-making where applicable.

Second, voice recordings constitute sensitive personal health data that require robust protection. Audio recordings can reveal identity, emotional state, and potentially clinical information beyond what was intended for collection. Data security frameworks must ensure encryption at rest and in transit, strict access controls, and compliance with applicable data protection regulations including the General Data Protection Regulation [63]. The anonymization procedures used in the present study, where the technical team received data without clinical identifiers, represent a minimum standard, but real-time clinical applications would require more comprehensive data governance.

Third, the consequences of classification errors must be carefully weighed. False positives, cases in which a patient with depression is incorrectly classified as a suicide attempter, could lead to unnecessary clinical interventions, increased patient anxiety, involuntary psychiatric holds, or erosion of therapeutic trust. False negatives, cases in which a recent suicide attempter is classified as a depression patient, carry potentially fatal consequences by providing false reassurance to clinicians. The asymmetric clinical consequences of these two error types should inform threshold selection, alert-generation strategies, and the design of future clinical decision-support systems. Because the present study did not evaluate prospective clinical deployment, no operating threshold for clinical use can be recommended based on the current data alone.

Fourth, the tool must be positioned unambiguously as a decision support system rather than an autonomous diagnostic instrument. Clinical judgment must remain the final arbiter of risk assessment and treatment decisions. The model's output, a probability estimate derived from acoustic features, provides one additional data point in a complex clinical evaluation. It does not replace the therapeutic relationship, clinical interview, or structured risk assessment. Clear communication to both clinicians and patients about what the system measures, what it does not measure, and how its outputs should be interpreted is essential to responsible implementation.

Finally, algorithmic transparency and external validation represent additional ethical requirements for AI-based clinical decision support. Models intended for suicide-risk assessment should be evaluated across independent populations, healthcare settings, and demographic groups to minimize unintended bias and ensure equitable performance before being considered for clinical implementation.

4.6. Limitations

Several limitations of the present study must be acknowledged. First, with 170,691 total parameters (170,307 trainable) and an effective independent sample size of 88 patients, the parameter-to-patient ratio is approximately 1940:1, which, while substantially reduced from previous architectural iterations, remains above thresholds associated with reliable generalization in medical machine learning. Larger training samples are generally recommended for reliable generalization in medical machine learning, and smaller studies tend to yield less stable and potentially optimistic performance estimates. Although the segmentation procedure yielded 8037 total spectrogram segments (6660 training, 1377 test), these segments are not independent: they share speaker identity, recording conditions, and 50% temporal overlap, making the patient count the appropriate denominator for assessing overfitting risk. The evidence of overfitting observed in training (training accuracy reaching approximately 98% over 60 epochs) is consistent with this parameter-to-patient imbalance, and the classification accuracy reported here should be interpreted as preliminary.

Second, although repeated patient-level stratified five-fold cross-validation was performed to obtain a more robust estimate of model performance, all analyses were conducted

on the same cohort of 88 patients. Consequently, the reported results remain internally validated and require confirmation in completely independent external cohorts before any conclusions regarding generalizability can be drawn.

Third, no external validation cohort was available. All analyses, including repeated cross-validation and subgroup analyses, were performed using the same study population. Although patient-level separation was strictly maintained throughout all analyses, validation on independent cohorts collected at different institutions remains essential before clinical translation.

Fourth, the study population was drawn from two Romanian psychiatric hospitals, and the recordings were conducted in Romanian. While the theoretical basis of voice-based emotion detection, grounded in autonomic and laryngeal physiology that is shared across human populations, suggests potential cross-linguistic applicability, this must be empirically validated. Cross-dataset and cross-language performance drops have been consistently reported in the speech emotion recognition literature [24,37], and generalization to other languages, recording environments, and cultural contexts cannot be assumed without dedicated validation studies.

Fifth, although SpecAugment and class-weighted training were incorporated to improve robustness and partially compensate for class imbalance, the relatively small dataset limits the diversity of acoustic conditions represented during training. Additional augmentation strategies and larger multicenter datasets may further improve robustness and generalization.

Sixth, although early stopping and adaptive learning-rate reduction were implemented to reduce overfitting, the high parameter-to-patient ratio remains a challenge. Future studies should investigate additional regularization strategies together with substantially larger training datasets.

Seventh, medication data was not available for the study participants. Psychotropic medications can influence voice characteristics through effects on psychomotor activity, laryngeal muscle tone, and autonomic function. The absence of medication data prevents us from assessing whether medication differences between groups contributed to the observed classification performance. This confound must be addressed in future studies through systematic recording and statistical control of medication profiles.

Eighth, although additional female-only and repeated sex-matched analyses were performed, sex imbalance remains an important limitation because the male-only cohort contained only five patients with depression. Consequently, the influence of sex cannot yet be considered fully resolved and should be addressed prospectively through balanced recruitment.

Ninth, the recordings were obtained during semi-structured clinical interviews in hospital settings, which introduces potential environmental confounds (background noise, recording equipment variability) and interaction effects (interviewer behavior, rapport quality). Standardization of recording protocols across sites was maintained, but the naturalistic clinical context differs from the controlled laboratory conditions used in many speech emotion recognition benchmarks. Additionally, the audio was encoded at 128 kbps in MP3 format, a lossy compression codec that removes spectral information, particularly at higher frequencies, below the standard of lossless WAV formats typically used in acoustic research. While the mel-spectrogram representation focuses on lower-frequency bands where emotional information is concentrated, some loss of fine-grained spectral detail cannot be excluded.

Tenth, because the classification labels are derived from the clinical event (suicide attempt vs. no attempt), the model learns to distinguish acoustic correlates of this clinical

distinction rather than identifying a prospective risk signal. The extent to which these acoustic patterns would generalize to pre-attempt detection remains to be demonstrated.

Eleventh, although interviewer speech was completely removed before acoustic analysis through speaker diarization and manual verification, interviewer-related influences cannot be completely excluded because interviewers were aware of participants' clinical status and the interview topics differed between the two groups. Therefore, part of the observed discrimination may reflect differences in emotional elicitation rather than exclusively suicide-related neurophysiological alterations.

Twelfth, no voice-based psychiatric diagnostic or risk assessment tool has received regulatory clearance (FDA, CE marking, or equivalent) to date. The present work constitutes early-stage evidence that must progress through the established regulatory pathway, including prospective external validation, clinical performance studies, and compliance with applicable frameworks such as the EU Medical Device Regulation (2017/745) and the EU AI Act, before any clinical deployment can be considered.

4.7. Future Directions

The findings of the present study motivate several lines of future work. The immediate priority is a larger, multi-site validation study enrolling a minimum of 500 participants across classification categories, drawn from multiple institutions and, where feasible, multiple linguistic populations. Such a study should include independent external validation cohorts collected at multiple institutions, balanced demographic representation, standardized interview protocols, and prospective evaluation across different languages and recording environments. Although the present study employed repeated patient-level stratified five-fold cross-validation with multiple random seeds, future studies should evaluate the robustness of the proposed approach under true external validation conditions and report discrimination, calibration, and clinical utility according to current TRIPOD-AI recommendations [64].

Extension to a multi-level classification scheme is planned. The present binary design (suicide attempters vs. depression without ideation) represents a first step. A three-level scheme incorporating patients at intermediate risk, such as treated patients with prior attempts who are in a transitional vulnerability state, would provide a graded risk assessment more closely aligned with clinical decision-making. This extension will require larger samples, particularly for the intermediate category, to ensure adequate statistical power.

A prospective longitudinal design would add critical clinical evidence by tracking patients from the index hospitalization through the post-discharge period, recording daily voice samples via a smartphone application, and evaluating whether changes in the model-derived risk score precede clinical deterioration or repeat suicidal behavior. This design would test the tool's potential for dynamic risk monitoring rather than one-time classification.

Integration with existing clinical assessment instruments should be explored. Future studies should also evaluate whether voice-based biomarkers provide incremental predictive value beyond established clinical variables and existing structured assessment instruments using appropriate multivariable statistical models. A multimodal approach combining the voice biomarker with structured instruments such as the C-SSRS and clinical judgment may yield superior performance compared with any modality in isolation. The theoretical expectation is that voice analysis would add the most incremental value in patients who deny ideation on self-report instruments, precisely the scenario where clinical risk assessment is currently weakest.

An additional priority will be to investigate the biological mechanisms underlying the observed acoustic discrimination. Future studies combining voice analysis with standardized depression severity scales, medication records, physiological biomarkers, and neuroimaging could help determine which components of the learned acoustic representation reflect suicide-related neurophysiological alterations and which represent correlated clinical or demographic factors. Such multimodal studies would substantially improve the interpretability of AI-based voice biomarkers.

Finally, the ethical and regulatory dimensions of deploying a voice-based psychiatric monitoring tool require careful navigation. Informed consent protocols must account for the capacity fluctuations inherent in acute psychiatric illness. The tool must be positioned unambiguously as clinical decision support, never as autonomous diagnosis, with transparent communication to patients about what the system measures and how its outputs are used in their care.

5. Conclusions

This study presents preliminary evidence that a Multi-Scale CNN with residual connections and spatial attention, applied to mel-spectrograms derived from clinical interview recordings, can classify patients who recently attempted suicide from patients with severe depression without suicidal ideation. Repeated patient-level stratified five-fold cross-validation yielded a mean classification accuracy of $72.7 \pm 6.4\%$ and a mean ROC-AUC of 0.824 ± 0.054 , providing a substantially more robust estimate of model performance than the single train-test split reported in the initial version of the manuscript. Across repeated cross-validation, the model achieved a mean sensitivity of $88.3 \pm 9.5\%$ and a mean specificity of $41.3 \pm 25.0\%$. The relatively small cohort of 88 patients (59 suicide attempters and 29 patients with severe depression without suicidal ideation), together with the class imbalance and the absence of external validation, underscores the exploratory nature of these findings.

The clinical rationale for this approach rests on two converging lines of evidence. First, the dissimulation problem, the well-documented tendency of suicidal patients to deny their ideation during clinical assessment, fundamentally limits all self-report-based instruments, creating a critical gap in risk detection precisely where the stakes are highest. Second, as reported in Section 1, the post-discharge period following psychiatric hospitalization is characterized by suicide rates that can exceed 6000 per 100,000 person-years in the first days after discharge, and current clinical responses to this risk window are intermittent and dependent on patient disclosure. A voice-based monitoring system could address both challenges by providing continuous, non-invasive assessment through an autonomically mediated channel that the patient cannot fully suppress. However, as discussed, the present study establishes a necessary precondition for such an approach, that the voices of recent attempters are acoustically separable from those of depressed patients without suicidal ideation, rather than a sufficient demonstration of dissimulation-resistant detection.

The study has important limitations, principally the relatively small sample size of 88 patients, the high parameter-to-patient ratio, the class imbalance, the single-language and two-site design, the absence of external validation, and several remaining potential confounding factors, including medication exposure, differences in interview context, and sex distribution. Although repeated patient-level cross-validation substantially improved the robustness of performance estimation, confirmation in large independent multicenter cohorts remains essential before clinical implementation can be considered. Nevertheless, the convergence of neurophysiological evidence for involuntary modulation of voice production, the well-documented limitations of self-report-based suicide assessment, and the reproducible acoustic discrimination observed in the present study provide a scientific

cally plausible foundation for the continued development of voice-based biomarkers as a complementary tool for suicide risk assessment.

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

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committees of “Prof. Dr. Al. Obregia” Psychiatry Clinical Hospital, Bucharest, Romania (Approval no. 39420/approval date 7 December 2023) and “Elisabeta Doamna” Psychiatry Hospital, Galati, Romania (Approval no. 2/approval date 29 January 2024).

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

Data Availability Statement: The audio recordings and clinical data used in this study cannot be made publicly available due to the sensitive nature of the material and ethical obligations to protect participant privacy. The recordings contain identifiable voice data from psychiatric inpatients, and their release would violate the informed consent agreements and data protection requirements under which the study was approved. The mel-spectrogram feature matrices (with no reversible audio content), model architecture code, and training scripts may be made available upon reasonable request to the corresponding author, subject to a data use agreement and approval by the relevant ethics committees. Inquiries regarding data access should be directed to contact@vldtau.com.

Acknowledgments: Publication of this paper was supported by the University of Medicine and Pharmacy Carol Davila, through the institutional program “Publish not Perish”.

Conflicts of Interest: Authors Vlad Tau, Catalin Niculae, and George Florian Macarie were employed by Eden Health SRL, Areus Technology, and Clinica PSIHOGER, respectively, at the time of the study and the preparation of the manuscript. These affiliations were unrelated to the funding of the present research. The respective employers had no financial or commercial interest in its outcomes. These authors therefore declare that their professional affiliations did not constitute a potential or actual conflict of interest. Their specific contribution types in the present study are stated above. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. World Health Organization. Suicide. Available online: <https://www.who.int/news-room/fact-sheets/detail/suicide> (accessed on 25 March 2025).
2. Eurostat. 1 in 6 Deaths of Young People Due to Suicide in 2022. Available online: <https://ec.europa.eu/eurostat/web/products-eurostat-news/w/edn-20250910-1> (accessed on 10 September 2025).
3. Centers for Disease Control and Prevention. Suicide Data and Statistics. Available online: <https://www.cdc.gov/suicide/facts/index.html> (accessed on 26 March 2025).
4. Franklin, J.C.; Ribeiro, J.D.; Fox, K.R.; Bentley, K.H.; Kleiman, E.M.; Huang, X.; Musacchio, K.M.; Jaroszewski, A.C.; Chang, B.P.; Nock, M.K. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychol. Bull.* **2017**, *143*, 187–232. [CrossRef] [PubMed]
5. de la Torre-Luque, A.; Pemau, A.; Ayad-Ahmed, W.; Borges, G.; Fernandez-Sevillano, J.; Garrido-Torres, N.; Garrido-Sanchez, L.; Garriga, M.; Gonzalez-Ortega, I.; Gonzalez-Pinto, A.; et al. Risk of suicide attempt repetition after an index attempt: A systematic review and meta-analysis. *Gen. Hosp. Psychiatry* **2023**, *81*, 51–56. [CrossRef] [PubMed]

6. Carter, G.; Milner, A.; McGill, K.; Pirkis, J.; Kapur, N.; Spittal, M.J. Predicting suicidal behaviours using clinical instruments: Systematic review and meta-analysis of positive predictive values for risk scales. *Br. J. Psychiatry* **2017**, *210*, 387–395. [[CrossRef](#)] [[PubMed](#)]
7. Belsher, B.E.; Smolenski, D.J.; Pruitt, L.D.; Bush, N.E.; Beech, E.H.; Workman, D.E.; Morgan, R.L.; Evatt, D.P.; Tucker, J.; Skopp, N.A. Prediction models for suicide attempts and deaths: A systematic review and simulation. *JAMA Psychiatry* **2019**, *76*, 642–651. [[CrossRef](#)] [[PubMed](#)]
8. Seyedsalehi, A.; Bailey, J.; Ogonah, M.G.T.; Fanshawe, T.R.; Fazel, S. Prediction models for self-harm and suicide: A systematic review and critical appraisal. *BMC Med.* **2025**, *23*, 549. [[CrossRef](#)] [[PubMed](#)]
9. Pigoni, A.; Delvecchio, G.; Turtulici, N.; Madonna, D.; Pietrini, P.; Cecchetti, L.; Brambilla, P. Machine learning and the prediction of suicide in psychiatric populations: A systematic review. *Transl. Psychiatry* **2024**, *14*, 140. [[CrossRef](#)] [[PubMed](#)]
10. Spittal, M.J.; Guo, X.A.; Kang, L.; Kirtley, O.J.; Clapperton, A.; Borschmann, R.; Pirkis, J.; Shin, J.I. Machine learning algorithms and their predictive accuracy for suicide and self-harm: Systematic review and meta-analysis. *PLoS Med.* **2025**, *22*, e1004581. [[CrossRef](#)] [[PubMed](#)]
11. Teismann, T.; Janssen, W.C.; Heering, H.D. Suicide risk assessment: Clinical implications of the unpredictability of suicidal behavior. *Front. Psychiatry* **2026**, *17*, 1844322. [[CrossRef](#)] [[PubMed](#)]
12. Large, M.M.; Ryan, C.J. Suicide risk categorisation of psychiatric inpatients: What it might mean and why it is of no use. *Australas. Psychiatry* **2014**, *22*, 390–392. [[CrossRef](#)] [[PubMed](#)]
13. Baston, R. Beyond prediction: A new paradigm for understanding suicide risk. *Synthese* **2024**, *204*, 35. [[CrossRef](#)]
14. Obegi, J.H. How common is recent denial of suicidal ideation among ideators, attempters, and suicide decedents? A literature review. *Gen. Hosp. Psychiatry* **2021**, *72*, 92–95. [[CrossRef](#)] [[PubMed](#)]
15. Berman, A.L. Risk factors proximate to suicide and suicide risk assessment in the context of denied suicide ideation. *Suicide Life-Threat. Behav.* **2018**, *48*, 340–352. [[CrossRef](#)] [[PubMed](#)]
16. Busch, K.A.; Fawcett, J.; Jacobs, D.G. Clinical correlates of inpatient suicide. *J. Clin. Psychiatry* **2003**, *64*, 14–19. [[CrossRef](#)] [[PubMed](#)]
17. Leenaars, A.A.; Dieserud, G.; Wenckstern, S. The mask of suicide. *Arch. Suicide Res.* **2022**, *26*, 1072–1093. [[CrossRef](#)] [[PubMed](#)]
18. Chung, D.T.; Ryan, C.J.; Hadzi-Pavlovic, D.; Singh, S.P.; Stanton, C.; Large, M.M. Suicide rates after discharge from psychiatric facilities: A systematic review and meta-analysis. *JAMA Psychiatry* **2017**, *74*, 694–702. [[CrossRef](#)] [[PubMed](#)]
19. Aaltonen, K.; Sund, R.; Hakulinen, C.; Pirkola, S.; Isometsa, E. Variations in suicide risk and risk factors after hospitalization for depression in Finland, 1996–2017. *JAMA Psychiatry* **2024**, *81*, 506–515. [[CrossRef](#)] [[PubMed](#)]
20. Chung, D.T.; Hadzi-Pavlovic, D.; Wang, M.; Swaraj, S.; Olfson, M.; Large, M. Meta-analysis of suicide rates in the first week and the first month after psychiatric hospitalisation. *BMJ Open* **2019**, *9*, e023883. [[CrossRef](#)] [[PubMed](#)]
21. Forte, A.; Buscajoni, A.; Fiorillo, A.; Pompili, M.; Baldessarini, R.J. Suicidal risk following hospital discharge: A review. *Harv. Rev. Psychiatry* **2019**, *27*, 209–216. [[CrossRef](#)] [[PubMed](#)]
22. Van Puyvelde, M.; Neyt, X.; McGlone, F.; Pattyn, N. Voice stress analysis: A new framework for voice and effort in human performance. *Front. Psychol.* **2018**, *9*, 1994. [[CrossRef](#)] [[PubMed](#)]
23. Iyer, R.; Meyer, D. Detection of suicide risk using vocal characteristics: Systematic review. *JMIR Biomed. Eng.* **2022**, *7*, e42386. [[CrossRef](#)] [[PubMed](#)]
24. Marie, A.; Garnier, M.; Bertin, T.; Machart, L.; Dardenne, G.; Quellec, G.; Berrouguet, S. Acoustic and machine learning methods for speech-based suicide risk assessment: A systematic review. *arXiv* **2025**, arXiv:2505.18195. [[CrossRef](#)]
25. Min, S.; Shin, D.; Rhee, S.J.; Park, C.H.K.; Yang, J.H.; Song, Y.; Kim, M.J.; Kim, K.; Cho, W.I.; Kwon, O.C.; 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. [[CrossRef](#)] [[PubMed](#)]
26. Su, Z.; Jiang, H.; Yang, Y.; Hou, X.; Su, Y.; Yang, L. Acoustic features for identifying suicide risk in crisis hotline callers: Machine learning approach. *J. Med. Internet Res.* **2025**, *27*, e67772. [[CrossRef](#)] [[PubMed](#)]
27. Pillai, A.; Nepal, S.; Wang, W.; Nemesure, M.; Heinz, M.; Price, G.; Lekkas, D.; Collins, A.C.; Griffin, T.; Buck, B.; et al. Investigating generalizability of speech-based suicidal ideation detection using mobile phones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2023**, *7*, 174. [[CrossRef](#)]
28. Lejeune, A.; Le Glaz, A.; Perron, P.-A.; Sebt, J.; Baca-Garcia, E.; Walter, M.; Berrouguet, S. Artificial intelligence and suicide prevention: A systematic review. *Eur. Psychiatry* **2022**, *65*, e19. [[CrossRef](#)] [[PubMed](#)]
29. Kusuma, K.; Larsen, M.; Quiroz, J.C.; Gillies, M.; Burnett, A.; Qian, J.; Torok, M. The performance of machine learning models in predicting suicidal ideation, attempts, and deaths: A meta-analysis and systematic review. *J. Psychiatr. Res.* **2022**, *155*, 579–588. [[CrossRef](#)] [[PubMed](#)]
30. Ehtemam, H.; Sadeghi Esfahlani, S.; Sanaei, A.; Ghaemi, M.M.; Hajesmaeel-Gohari, S.; Rahimisadegh, R.; Bahaadinbeigy, K.; Ghasemian, F.; Shirvani, H. Role of machine learning algorithms in suicide risk prediction: A systematic review-meta analysis of clinical studies. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 138. [[CrossRef](#)] [[PubMed](#)]

31. Fernandez-Quijano, I.; Herrera-Peco, I.; Lopez-Espuela, F.; Suarez-Llevat, C.; Moreno-Sanchez, R.; Ruiz-Nunez, C. Artificial intelligence in suicide prevention: A systematic review of randomized controlled trials on risk prediction, fully automated interventions, and AI-guided treatment allocation. *Psychiatry Int.* **2025**, *6*, 143. [\[CrossRef\]](#)
32. Liu, L.; Li, Z.; Hu, Y.; Li, C.; He, S.; Zhang, S.; Gao, J.; Zhu, H.; Huang, G. Predictive performance of machine learning for suicide in adolescents: Systematic review and meta-analysis. *J. Med. Internet Res.* **2025**, *27*, e73052. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Buscher, R.; Winkler, T.; Mocellin, J.; Homan, S.; Josifovski, N.; Ciharova, M.; van Breda, W.; Kwon, S.; Larsen, M.E.; Torous, J.; et al. A systematic review on passive sensing for the prediction of suicidal thoughts and behaviors. *npj Ment. Health Res.* **2024**, *3*, 42. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Quellec, G.; Berrouguet, S.; Morgieue, M.; Dubois, J.; Leboyer, M.; Vaiva, G.; Aze, J.; Courtet, P. Predicting suicidal ideation from irregular and incomplete time series of questionnaires in a smartphone-based suicide prevention platform: A pilot study. *Sci. Rep.* **2024**, *14*, 20870. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Grimland, M.; Liberman, M.; Yeshayahu, H.; Benatov, J.; Munz, N.; Segal, A.; Ben Dayan, L.; Shenfeld, I.; Gal, K.; Levi-Belz, Y. Explainable AI for suicide risk detection: Gender- and age-specific patterns from real-time crisis chats. *Front. Med.* **2025**, *12*, 1703755. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Amiriparian, S.; Gerczuk, M.; Lutz, J.; Strube, W.; Papazova, I.; Hasan, A.; Kathan, A.; Schuller, B.W. Non-invasive suicide risk prediction through speech analysis. *arXiv* **2024**, arXiv:2404.12132. [\[CrossRef\]](#)
37. Krautz, A.E.; Volkening, J.; Raue, J.; Otte, C.; Eickhoff, S.B.; Ahlers, E.; Langner, J. Prediction of suicide using web based voice recordings analyzed by artificial intelligence. *Sci. Rep.* **2025**, *15*, 23855. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Franti, E.; Ispas, I.; Dragomir, V.; Dascalu, M.; Elteto, Z.; Stoica, I.C. Voice based emotion recognition with convolutional neural networks for companion robots. *Rom. J. Inf. Sci. Technol.* **2017**, *20*, 222–240.
39. Franti, E.; Dascalu, M.; Ispas, I.; Tebeanu, A.V.; Elteto, Z.; Branea, S.; Dragomir, V. Decoding communication: A deep learning approach to voice-based intention detection. *Rom. J. Inf. Sci. Technol.* **2018**, *21*, 460–474.
40. Singh, Y.B.; Goel, S. A systematic literature review of speech emotion recognition approaches. *Neurocomputing* **2022**, *492*, 245–263. [\[CrossRef\]](#)
41. Lieskovska, E.; Jakubec, M.; Jarina, R.; Chmulik, M. A review on speech emotion recognition using deep learning and attention mechanism. *Electronics* **2021**, *10*, 1163. [\[CrossRef\]](#)
42. Mihalache, S.; Burileanu, D.; Franti, E.; Dascalu, M.; Bratan, C.-A. Lasting emotions—An investigation of short- and long-term affective content remanence in speech. *Rom. J. Inf. Sci. Technol.* **2022**, *25*, 20–35.
43. Bratan, C.A.; Tebeanu, A.V.; Bobes, G.; Popescu, I.; Iorgulescu, G.; Neagu, L.; Apostol, A.; Goian, R.; Dascalu, M.; Franti, E.; et al. Using swear words increases the irritability—A study using AI algorithms. *Rom. J. Inf. Sci. Technol.* **2023**, *26*, 365–374. [\[CrossRef\]](#)
44. Franti, E.; Ispas, I.; Dascalu, M. Testing the universal baby language hypothesis—Automatic infant speech recognition with CNNs. In Proceedings of the 41st International Conference on Telecommunications and Signal Processing (TSP), Athens, Greece, 4–6 July 2018; pp. 424–428. [\[CrossRef\]](#)
45. Bratan, C.A.; Gheorghe, M.; Ispas, I.; Franti, E.; Dascalu, M.; Stoicescu, S.M.; Rosca, I.; Gherghiceanu, F.; Dumitrache, D.; Nastase, L. Dunstan baby language classification with CNN. In Proceedings of the 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 13–15 October 2021; pp. 167–171. [\[CrossRef\]](#)
46. Wu, W.; Cui, Z.; Lei, C.; Duan, Y.; Qu, D.; Wu, J.; Zhou, B.; Chen, R.; Zhang, C. The 1st SpeechWellness challenge: Detecting suicide risk among adolescents. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025. [\[CrossRef\]](#)
47. Martin, V.P.; Brazier, C.; Amblard, M.; Musiol, M.; Rouas, J.-L. Network of acoustic characteristics for the automatic detection of suicide risk from speech. Contribution to the 2025 SpeechWellness challenge by the Semawave team. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025. [\[CrossRef\]](#)
48. France, D.J.; Shiavi, R.G.; Silverman, S.; Silverman, M.; Wilkes, D.M. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Trans. Biomed. Eng.* **2000**, *47*, 829–837. [\[CrossRef\]](#) [\[PubMed\]](#)
49. Ozdas, A.; Shiavi, R.G.; Wilkes, D.M.; Silverman, M.K.; Silverman, S.E. Analysis of vocal tract characteristics for near-term suicidal risk assessment. *Methods Inf. Med.* **2004**, *43*, 36–38. [\[CrossRef\]](#)
50. Ozdas, A.; Shiavi, R.G.; Silverman, S.E.; Silverman, M.K.; Wilkes, D.M. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Trans. Biomed. Eng.* **2004**, *51*, 1530–1540. [\[CrossRef\]](#) [\[PubMed\]](#)
51. Belouali, A.; Gupta, S.; Sourirajan, V.; Yu, J.; Allen, N.; Alaoui, A.; Dutton, M.A.; Reinhard, M.J. Acoustic and language analysis of speech for suicidal ideation among US veterans. *BioData Min.* **2021**, *14*, 11. [\[CrossRef\]](#) [\[PubMed\]](#)
52. Stasak, B.; Epps, J.; Schatten, H.T.; Miller, I.W.; Provost, E.M.; Arme, M.F. Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt. *Speech Commun.* **2021**, *132*, 10–20. [\[CrossRef\]](#) [\[PubMed\]](#)
53. Helou, L.B.; Jennings, J.R.; Rosen, C.A.; Wang, W.; Verdolini Abbott, K. Intrinsic laryngeal muscle response to a public speech preparation stressor: Personality and autonomic predictors. *J. Speech Lang. Hear. Res.* **2020**, *63*, 2940–2951. [\[CrossRef\]](#) [\[PubMed\]](#)
54. DeVille, D.C.; Kerr, K.L.; Avery, J.A.; Burrows, K.; Bodurka, J.; Feinstein, J.; Khalsa, S.S.; Paulus, M.P.; Simmons, W.K. Diminished responses to bodily threat and blunted interoception in suicide attempters. *eLife* **2020**, *9*, e51593. [\[CrossRef\]](#) [\[PubMed\]](#)

55. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech emotion recognition from spectrograms with deep convolutional neural network. In Proceedings of the International Conference on Platform Technology and Service (PlatCon), Busan, Republic of Korea, 13–15 February 2017; pp. 1–5. [\[CrossRef\]](#)
56. Horowitz, L.M.; Bridge, J.A.; Teach, S.J.; Ballard, E.; Klima, J.; Rosenstein, D.L.; Wharff, E.A.; Ginnis, K.; Cannon, E.; Joshi, P.; et al. Ask Suicide-Screening Questions (ASQ): A brief instrument for the pediatric emergency department. *Arch. Pediatr. Adolesc. Med.* **2012**, *166*, 1170–1176. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Olfson, M.; Wall, M.; Wang, S.; Crystal, S.; Liu, S.M.; Gerhard, T.; Blanco, C. Short-term suicide risk after psychiatric hospital discharge. *JAMA Psychiatry* **2016**, *73*, 1119–1126. [\[CrossRef\]](#) [\[PubMed\]](#)
58. Tay, J.L.; Li, Z. Brief contact interventions to reduce suicide among discharged patients with mental health disorders—A meta-analysis of RCTs. *Suicide Life-Threat. Behav.* **2022**, *52*, 1074–1095. [\[CrossRef\]](#) [\[PubMed\]](#)
59. Larsen, E.; Murton, O.; Song, X.; Joachim, D.; Watts, D.; Kapczinski, F.; Venesky, L.; Hurowitz, G. Validating the efficacy and value proposition of mental fitness vocal biomarkers in a psychiatric population: Prospective cohort study. *Front. Psychiatry* **2024**, *15*, 1342835. [\[CrossRef\]](#) [\[PubMed\]](#)
60. Schaffer, A.; Sinyor, M.; Kurdyak, P.; Vigod, S.; Sareen, J.; Reis, C.; Green, D.; Bolton, J.; Grigoriadis, S.; Cairney, J.; et al. Population-based analysis of health care contacts among suicide decedents: Identifying opportunities for more targeted suicide prevention strategies. *World Psychiatry* **2016**, *15*, 135–145. [\[CrossRef\]](#) [\[PubMed\]](#)
61. O'Connor, L.; Larkin, C.; Ibrahim, A.F.; Allen, M.; Wang, B.; Boudreaux, E.D. Development and pilot study of simple suicide risk rulers for use in the emergency department. *Gen. Hosp. Psychiatry* **2018**, *63*, 97–102. [\[CrossRef\]](#) [\[PubMed\]](#)
62. Mullinax, S.; Chalmers, C.E.; Brennan, J.; Vilke, G.M.; Nordstrom, K.; Wilson, M.P. Suicide screening scales may not adequately predict disposition of suicidal patients from the emergency department. *Am. J. Emerg. Med.* **2018**, *36*, 1779–1783. [\[CrossRef\]](#) [\[PubMed\]](#)
63. General Data Protection Regulation (GDPR). Available online: <https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=celex:32016R0679> (accessed on 1 July 2026).
64. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.