Artificial intelligence for suicide assessment using Audiovisual Cues: a review

Artificial Intelligence for Suicide Assessment using Audiovisual Cues: A Review

Sahraoui Dhelim [1], Liming Chen [2], Huansheng Ning [3*] and Chris Nugent [2]

[1]School of Computer Science, Dublin College University, Ireland
[2]School of Computing, Ulster University, United Kingdom
[3] School of Computer and Communication Engineering, University of Science and Technology Beijing, China
[*]Corresponding author: ninghuansheng@ustb.edu.cn

Abstract

Death by suicide is the seventh leading death cause worldwide. The recent advancement in Artificial Intelligence (AI), specifically AI applications in image and voice processing, has created a promising opportunity to revolutionize suicide risk assessment. Subsequently, we have witnessed fast-growing literature of research that applies AI to extract audiovisual non-verbal cues for mental illness assessment. However, the majority of the recent works focus on depression, despite the evident difference between depression symptoms and suicidal behavior non-verbal cues. In this paper, we review the recent works that study suicide ideation and suicide behavior detection through audiovisual feature analysis, mainly suicidal voice/speech acoustic features analysis and suicidal visual cues. Automatic suicide assessment is a promising research direction that is still in the early stages. Accordingly, there is a lack of large datasets that can be used to train machine leaning and deep learning models proven to be effective in other, similar tasks.

Keywords: suicide detection; machine learning; speech analysis; visual cues; suicide ideation detection;

1. Introduction

Suicide is a global mental health problem (World Health Organization 2014). More than 800,000 people pass away from suicide, and more than 16 million people attempt suicide every year (World Health Organization 2014). Suicide is currently the fourth leading cause of death for people aged between 15 and 29 years (World Health Organization 2019). It is still a challenge to assess suicide risks and detect suicide due to the transient and ambivalent nature of severe suicide intent, and the patient's hesitancy. Most suicide risk assessment methods depend on voluntary disclosure of the patient, and most suicidal patients deny suicide ideation during an interview (Blanchard and Farber 2020)(Warner et al. 2011). In such cases, the clinician relies on various secondary information sources, such as the patient's health records, the nurse's observation, and other secondary factors. However, the final judgment about the suicide risk level is based on the clinician's intuition and observation of the patient's behaviors during the interview. Although the clinician can observe obvious emotions and body gestures, the clinician cannot observe micro facial expressions and biosignals such as heartbeat rate and brain waves. Therefore, building an Artificial Intelligence (AI) enabled suicide risk assessment tool that observes the patient's biomarker and audiovisual cues and learns suicide ideation patterns might be a decisive indicator that could help the clinician to reach a clear judgment (Wang et al. 2021a). With the recent advances in the field of user-centered computing (Dhelim et al. 2020, 2022), automatic mental disorders detection using audiovisual data has become an active research topic. The patient's acoustic and visual markers are analysed to detect mental disorders such as depression (Pampouchidou et al. 2017). The recent advancement in AI, specifically AI applications in image, voice processing and natural language processing (He et al. 2022; Zhang et al. 2022), has created a promising opportunity to revolutionize suicide risk assessment. Subsequently, we have witnessed fast-growing literature of research that applies AI to extract audiovisual non-verbal cues for mental illness assessment (Castillo-Sánchez et al. 2020). Although that previous research has found associations between acoustic features and suicide tendency (Cummins et al. 2015)(Scherer et al. 2013); suicidal patients have

lower acoustic energy, breathy voice quality, and abnormal glottal control compared to people without suicide ideation history (Ozdas et al. 2004a). However, the majority of the recent works focus on depression (Ji et al. 2020), despite the evident difference between depression signs and suicidal behavior non-verbal cues. In this paper, we review the recent works that study suicide ideation and suicide behavior detection through audiovisual feature analysis, mainly suicidal voice/speech acoustic features analysis and suicidal visual cues analysis. Suicidal ideation can vary in presentation and severity, but generally it can be classified as one of the following phases (Liu et al. 2020). (1) passive suicide ideation: when the subject has thoughts of suicide or self-harm but no plan to carry it out. (2) active suicide ideation: when the subject has thoughts of engaging in suicide-related behavior and has suicidal intent and/or had developed a plan to carry it out. (3) suicidal attempt: the subject had attempted suicide and/or still attempting to suicide following an unsuccessful suicide attempt (Silverman et al. 2007).

The focus of the current review is automatic suicide ideation and suicide behavior detection through audiovisual feature analysis, with a special focus on works that used machine learning and deep learning approaches. We limit the coverage of the review to works published between 2004 and 2021. We have used PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework (Moher et al. 2010) guidelines to select publications related to automatic suicidal behavior detection. As shown in Figure 1, initially, 251 related papers between January 2004 and November 2021 were identified after searching Google Scholar, Elsevier, Web of Science, ACM Digital Library, IEEE Xplore digital library, Springer and PubMed for articles related to the following research queries: “automatic suicide detection”, “machine learning suicide detection”, “suicidal speech analysis”, “suicidal visual cues”, “suicide ideation detection”. The searches were limited to articles written in English. 455 additional articles were identified as related works; these papers were found by tracking the citation map of the searched articles. After removing duplicated articles, a total of 706 papers were collected in the identification phase. In the screening phase, based on the title and abstract screening 551 papers were excluded for not meeting the inclusion criteria. The majority of these papers either studied suicidal behavior from a clinical perspective, without automatic detection, or they use automatic detection for detecting mental disorders such as anxiety and minor depression. 104 articles were excluded in the eligibility phase after full-text reading. Finally, 51 articles were qualified for final inclusion.

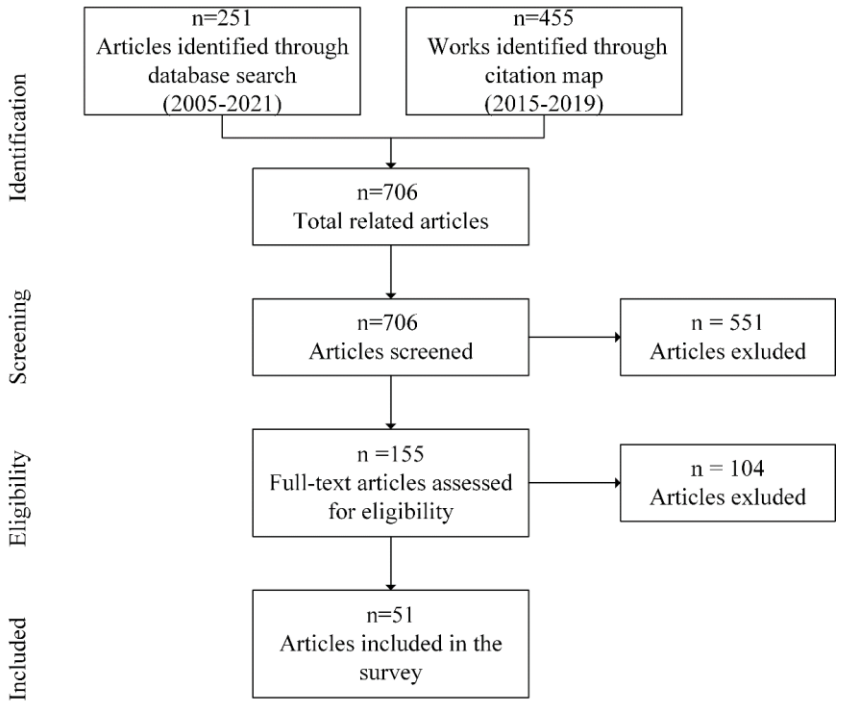

Figure 1 PRISMA flowchart of the review phases

The suicidal marker that can be extracted from audiovisual data, can be roughly divided into two main categories. Audio markers that convey the suicidal cue can be obtained by analyzing the acoustic features of the speech, and visual markers that can be observed in the suicidal patient's upper body behaviors during suicidal assessment procedures such as a clinical interview. Figure 2 shows the most used acoustic and visual features used in the literature on suicidal assessment using audiovisual data.

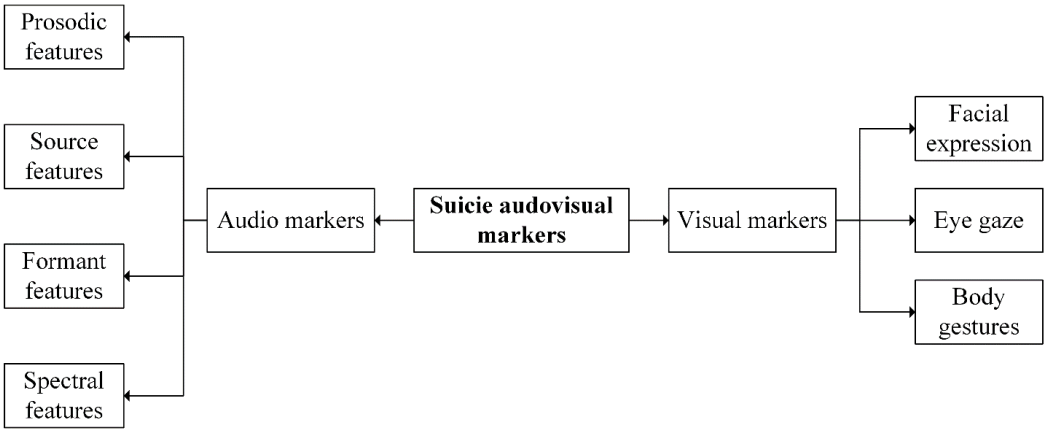

Figure 2 Suicidal audiovisual markers classification

The rest of the paper is organized as follows:

In Section 2, we discuss data collection scenarios of the reviewed works and summarize some of the most used audiovisual datasets for automatic suicide assessment. In Section 3, we focus on speech analysis and

audio-based suicide assessment. In Section 4, we review the works that used upper body features extracted from visual cues for suicide assessment. While Section 5 focuses on the detection models applied to audiovisual data for suicide behavior detection. In Section 6, we summarize the findings and outline limitations and potential future directions. Finally, Section 7 concludes the review.

2. Data collection

Suicide audiovisual data can be collected using three methods. (1) through an open interview between the patient and the clinician; (2) by capturing the audiovisual recording during casual conversations; (3) or by asking patients to read a speech or a set of sentences. Each one of these methods has its advantages and drawbacks. Because the recorded sentence is the same for all participants, the reading task method makes it easy to distinguish a suicidal patient from non-suicidal controls, which enable the machine learning model to focus on the audiovisual feature differences between suicidal patients and the control participants. The second advantage of the reading task method is that the model requires fewer data to learn, as it does not need to learn the context of the spoken sentence from the data, unlike the interview method, where the model needs to deduce the context of the spoken sentences.

In the interview method, unlike the reading task, the reason behind sudden changes in one of the acoustic markers must be distinguished from the context of the interview. For example, is the detected change due to grammatical contexts, such as intonation, or a suicidal cue? The second advantage of the reading task is that the clinician can design the reading passage to target specific sounds or syllables that are related to certain suicidal cue patterns. One of the famous reading tasks is known as the "rainbow passage", which is used in speech science because it contains all of the normal sounds in spoken English, and is phonetically balanced (Rohlfing et al. 2021). Table 1 presents study samples and data collection scenarios of some reviewed works. Whereas Table 2 summarizes some of the most used datasets that contain audiovisual recordings along with ground truth label regarding the mental conditions of the recoded individuals.

Table 1 Studied samples and data collection scenarios

Ref	N	Sample	Collection scenario
(Chakravarthula et al. 2020)	124	military couples	Recorded home conversations of veteran couples, following that they subjects were asked to take reasons-for-living and relationship counselling interviews. The interviews are 10 minutes long where the subjects were asked to discuss about life meaning and the their relationship conflicts.
(Venek et al. 2017)(Venek et al. 2014)	60	Adolescent	Total of 30 male and 30 females were asked to fill Suicidal Ideation Questionnaire-Junior (SIQ-JR) and Columbia Suicide Severity Rating Scale (C-SSRS), and take an interview with a trained social worker.
(Belouali et al. 2021)	588	Army veterans	Subjects regularly complete suicide assessment questionnaire in audio, and submit the recording via an Android app. The questionnaires contains open-ended questions, where the subjects can express their views about life meaning and other suicide related topics.
(Nasir et al. 2017)	54	military personnel	The subjects were interviewed for 10 minutes to 1 hour by 5 therapists specialized in suicide risk assessment, where the

			subjects answered Beck Scale for Suicidal Ideation (BSSI) and the Suicide Attempt Self-Injury Interview (SASII) questionnaires. The interview conversations were recorded using microphone in controlled environment.
(Galatzer-Levy et al. 2021)	20	Psychiatric hospital patients with suicidal history	Collected interview video recording of suicidal patients in hospital, and remote patient monitoring video using smartphones app when these patients were discharged from the hospital.
(Gideon et al. 2019)	43	Suicidal patients	Following suicide patients discharge from hospital, they were asked to install android app that records their daily phone conversations. A total of 4,078 calls over 402 hours were collected.
(Stasak et al. 2021)	246	Suicidal patients	Collected audio recordings of suicidal patients reading set of selected sentences, such as "my life has meaning".
(Laksana et al. 2017)	379	126 mentally ill patients, and 130 suicidal patients	Collected the video recording of every patients when answering interview questions for 8 minutes.
(Shah et al. 2019)	90	Suicidal social media users	Collected social media videos that were shared by suicidal patients. And only the videos where the subjects were talking directly to the camera is selected.
(Kleiman and Rule 2013)	40	Suicidal students	Extracted photos of suicidal students from their high school yearbook in the past.
(Eigbe et al. 2018)	379	126 mentally ill patients, and 130 suicidal patients	Collected the video recording of every patient when answering interview questions for 8 minutes.
(Pestian et al. 2018)	253	Suicidal patients	Recorded the subjects' responses to standardized suicide interviews designed to harvest thought markers.
(Scherer et al. 2013)	60	teenagers between the ages of 12 and 17	Collected the video recording of subjects when answering 5 open-ended questions related to suicide ideation. The recordings were collected in a controlled examination room using a tabletop microphone.
(France et al. 2000)	22	Suicidal patients	Collected the audio of suicidal patients when interviewed by therapists.
(Yingthawornsuk et al. 2006)	23	Suicidal patients	Collected the audio recording of each patient during the interview with the therapist, as well as when the patient is reading a pre-selected rainbow passage from a book.

(Ozdas et al. 2004a)	20	Suicidal patients	Collected the audio recording of suicidal patients during real-life situations, the recording was filtered and only high-quality recordings were considered.
(Akkaralaertsest and Yingthawornsuk 2015)	30	Suicidal and depressed patients	Collected the audio recording of each patient during the interview with the therapist, as well as when the patient is reading a pre-selected rainbow passage from a book.
(Anunvrapong and Yingthawornsuk 2014)	10	Suicidal and depressed patients	Collected the audio recording of the subject's speech when answering interview questions in a controlled environment.
(Yingthawornsuk and Shiavi 2008)	30	Suicidal and depressed patients	Collected the audio recording of each patient during the interview with the therapist, as well as when the patient is reading a pre-selected rainbow passage from a book.
(Ozdas et al. 2004b)	30	Suicidal and depressed patients	Collected the audio recording of the subject's speech when answering interview questions in a controlled environment.
(Keskinpala et al. 2007)	30	Suicidal and depressed patients	Collected the audio recording of the patient reading a pre-selected rainbow passage from a book. The rainbow passage is selected to includes all the sounds of spoken English, and phonetically balanced.
(NH et al. 2015b)(NH et al. 2015a)	23	Suicidal and depressed patients	The patients were recorded during the interview in a soundproof acoustically ideal room, where the patients were asked to take Hamilton Depression Rating Scale (HAMD) and Beck Depression Inventory (BDI-II). Following that, the patients were asked to read a rainbow passage.
(Yingthawornsuk et al. 2007)	20	Suicidal and depressed patients	One is a speech sample recorded from a clinical interview with a therapist and another is a speech sample recorded from a text–reading session.

Table 2 Audio-visual datasets for mental health research.

Dataset	Description	Language	Access
DIAC-WOZ (Gratch et al. 2014)	Contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. Data collected include audio and video recordings and extensive questionnaire responses; this part of the corpus includes data from the Wizard-of-Oz interviews, conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room.	English	Private

AVDLC (Valstar et al. 2013)	Audio-visual depressive language corpus (AViD-Corpus), includes 340 video clips of subjects performing a Human-Computer Interaction task while being recorded by a webcam and a microphone.	German, English	Public
AVEC 2014 (Ringeval et al. 2019)	The Audio/Visual Emotion Challenge 2014 (AVEC) is a subset of AVDLC dataset.	German, English	Public

3. Suicidal audio markers

Audio markers are classified as prosodic features, source features, formant features, and spectral features.

3.1. Source features

Source features are the main influencers of the changes in voice quality, changes in vocal fold vibration, loudness, vocal tract shape and pitch variation.

3.1.1. Source features and suicidal

Source features measure the quality of the voice production process and observe the changes as the airflow from the lungs passes through the glottis. Table 3 lists the most used source features in automatic suicide assessment. When the patient is in a suicidal state, the laryngeal control is affected, such as harshness, breathiness and creakiness, hence this change is reflected in the source feature (Drugman et al. 2012). The source features are either extracted by measuring the changes of glottal characteristics or changes in vocal fold movements and subsequently the change in voice quality features. The majority of source features are computed by analyzing the time-series data of the glottal flow signal (Doval et al. 2006; Airas 2008). Nonetheless, it is challenging to determine the critical time points because of the non-uniform vocal fold activity, and signal noise (Walker and Murphy 2007). Source features widely used in suicidal speech analysis include the jitter, which represents the cycle-to-cycle fluctuations in glottal pulse timing during a suicidal speech; shimmer, which represents the cycle-to-cycle fluctuations in glottal pulse amplitude in voiced regions; and harmonic-to-noise ratio (HNR), which represent the ratio of harmonics to inharmonic. These source features have been proven to be correlated with suicidality (Ozdas et al. 2004a)(Quatieri and Malyska 2012), that is because they are directly related to vocal fold vibration, which influences the vocal fold tension and subglottal pressure (Sundberg et al. 2011).

Table 3 Source feature used in automatic suicide assessment.

Feature	Description	Study	Significant test
Jitter	Deviations in individual consecutive f0 period lengths, which indicates irregular closure and asymmetric vocal-fold vibrations	(Chakravarthula et al. 2020) (Belouali et al. 2021) (Gideon et al. 2019)(Ozdas et al. 2004a)	Control: 0.0165 ± 0.002 (n = 10) Suicide: 0.0217 ± 0.005 (n = 10) $p \leq 0.05$ (t-test)
Shimmer	The difference in the peak amplitudes of consecutive f0 periods, indicates irregularities in voice intensity.	(Chakravarthula et al. 2020) (Belouali et al.	$p \leq 0.05$

		2021)(Ozdas et al. 2004a) (Gideon et al. 2019)	
QOQ	The ratio of the vocal folds' opening time. Functional dysphonias often reduce QOQ range. Speaking loudly requires more effort with a low QOQ and sounds more stalled.	(Venek et al. 2017) (Belouali et al. 2021) (Venek et al. 2014) (Scherer et al. 2013)	Control: 0.31 ± 0.13 (n = 8) Suicide: 0.42 ± 0.2 (n = 8) $p \leq 0.002$ (t-test)
NAQ	The ratio between peak-to-peak pulse amplitude and the negative peak of the differentiated flow glottogram and normalized with respect to the period time. It can be an estimate of glottal adduction.	(Venek et al. 2017) (Belouali et al. 2021)(Venek et al. 2014) (Scherer et al. 2013)	Control: 0.09 ± 0.04 (n = 8) suicide: 0.12 ± 0.05 (n = 8) $p \leq 0.002$ (t-test)
Spectral slope	Measures the glottal flow waveform needed for the reconstruction of the excitation signal from the given speech signal by glottal inverse filtering. (Ozdas et al. 2004a)	(Gideon et al. 2019) (Ozdas et al. 2004a) (Scherer et al. 2013) (Akkaralaertsest and Yingthawornsuk 2015)	Control: −83.3 ± 5.46 (n = 10) Suicide: −75.56 ± 8.53 (n = 10) $p \leq 0.05$ (t-test)
Vocal tract	Measures the change in vocal tract dynamics and constrain articulatory movement	(Yingthawornsuk and Shiavi 2008) (Yingthawornsuk et al. 2007)	N/A

3.1.2 Source features analysis

Due to the difficulties of accurately extracting source features from speech signals, most of the works in automatic suicide detection fail to establish a strong correlation between source features only and suicide, therefore the source features are usually combined with features from other audio categories. Another influencing factor when extracting source features such as shimmer, jitter and HNR is the speech type from which these features were extracted, such as continuous speaking or held vowels context. Generally, the held vowels make the extraction of source features much easier, but they are prone to an error related to individual-related differences in sound pressure levels, which could lead to inaccuracy in feature analysis. Whereas analyzing continuous speech to extract source features is more challenging than extracting them from held vowels, as it is extremely difficult to automatically determine voiced segments in the studied utterance (Kane et al. 2014). Identification of speech sections that contain significant source features is an active research direction in the field of suicidal speech analysis (Kane et al. 2014). Chakravarthula et al.

(Chakravarthula et al. 2020) analyzed source acoustic features in military couples' conversations and tried to deduce suicidal markers from these acoustic markers. Specifically, they have combined voice quality features (jitter, shimmer) with features from other acoustic categories, mainly prosodic features, and spectral features. They fused these acoustic features with lexical features Linguistic Inquiry and Word Count (LIWC), and behavioral features and applied a support vector machine (SVM) to classify the subjects into various classes (risk vs no-risk classes, degree of risk classes, non-severe vs severe risk classes). Scherer et al. (Scherer et al. 2013) analyzed speech properties of prosodic as well as source features that were extracted from the dyadic interview corpus of both suicidal and non-suicidal adolescents. They have found various statistically significant differences between the speech properties of suicidal and healthy subjects. The source and voice quality features exhibit the strongest differences between the two subject groups. Specifically, NAQ, QOQ and peak features are strongly associated with voice qualities on the tense dimension to breathy, which show that suicidal subjects' voices are breathier than the voice of healthy people. Venek et al. (Venek et al. 2017) [7] fused source features (NAQ, QOQ) with prosodic features and formant features to analyse suicidal speech. They applied hierarchical classifier (SVM then AdaBoostM1) on speech of 60 audio-recorded dyadic clinician-patient interviews of 30 suicidal patients. They found that there were 37 statistically significant features that could distinguish suicidal from non-suicidal subjects, that including: Speech time, Pause time, Personal Pron, 1st person Pron, Impersonal Pron, Past Tense, Negation, Positive emotion, Negative emotion, Tentative, Death, Nonfluencie, Assent, NAQ and QOQ.

3.2. Prosodic features

Prosodic features measure longitudinal variations in a speaker's rhythm, stress level, and intonation of speech. Some of the prominent examples of prosodic features are the speaking rate, the speaker's pitch and loudness level, and the speaker's energy dynamics. Among these prosodic features, the energy, and the fundamental frequency (also known as f0, e.g. the rate of vocal fold vibration) and are the most commonly used prosodic features, as they are direct indicators of pitch and loudness.

3.2.1. Prosodic features and suicidal

Assessing depressed and suicidal patients based on prosodic features has been investigated in early paralinguistic studies. Depressed and suicidal patients showed prosodic speech abnormalities such as reduced pitch range, articulation errors, slower speaking rate and reduced pitch. One of the earliest efforts in this regard is the work of Dr. S. Silverman. After analyzing recorded psychiatry sessions of suicidal patients, he discovered that in pre-suicidal phase the patients' speech can be distinguished through noticeable changes in its quality (Ozdas et al. 2004a). Silverman noticed that speech production mechanisms of pre-suicidal patients have altered the acoustic features of speech in measurable ways (Silverman et al. 2006). Table 4 lists the most used prosodic features in automatic suicide assessment.

Table 4 Prosodic feature used in automatic suicide assessment.

Feature	Description	Study	Significant test
f0	Fundamental frequency: lowest frequency of the speech signal, perceived as pitch (mean, median).	(Belouali et al. 2021)(Venek et al. 2014)(Gideon et al. 2019) (Scherer et al. 2013)(Fra	Control: 220.82 Suicide: 150.62 $p \leq 0.001$ (Belouali et al. 2021)

		nce et al. 2000)	
f0 variability	Measures of dispersion of f0 (variance, standard deviation).	(Venek et al. 2017)(Belouali et al. 2021)(Venek et al. 2014)(Gideon et al. 2019) (Scherer et al. 2013)(France et al. 2000)	N/A
f0 range	Difference between the lowest and highest f0 values.	(Belouali et al. 2021)(Venek et al. 2014)(Gideon et al. 2019) (Scherer et al. 2013) (France et al. 2000)	N/A
Intensity mean	Defined as the acoustic intensity (i.e., power carried by sound per unit area in a direction perpendicular to that area) in decibels relative to a reference value, perceived as loudness.	(Chakravarthula et al. 2020)	N/A
Intensity variability	Measures of dispersion of intensity (variance, standard deviation).	(Chakravarthula et al. 2020)	N/A
Energy	Measured as the mean-squared central difference across frames and may correlate with motor coordination	(Belouali et al. 2021) (Nasir et al. 2017)(Gid	Control: 2.8806 (n=504) Suicide: 2.3614 (n=84) $p \leq 0.001$ (Belouali et al. 2021)

		eon et al. 2019) (Scherer et al. 2013) (Keskinpala et al. 2007)	
Maximum phonation time	The mean of three attempts of the following measure is taken: the maximum time during which phonation of a vowel (usually lal) is sustained as long as possible with an upright position, deep breath, and a comfortable pitch and loudness	(Belouali et al. 2021)	N/A
Speech rate	Number of speech utterances per second over the duration of the speech sample (including pauses).	(Venek et al. 2017)	N/A
Time talking	Sum of the duration of all speech segments.	(Venek et al. 2017)	N/A
Pause duration mean	Mean duration of pause length.	(Venek et al. 2017)	N/A
Pause variability	Measures of dispersion of pause duration (variance, standard deviation).	(Venek et al. 2017)	N/A
Pause rate	Total length of pauses divided by the total length of speech (including pauses).	(Venek et al. 2017)	N/A
Pauses total	Total duration of pauses.	(Venek et al. 2017)	N/A
Pitch	The pitch level	(Chakravarthula et al. 2020) (Belouali et al. 2021)(Nasir et al. 2017) (Gideon et al. 2019) (Shah et al. 2019)	N/A
PSP	Measure is derived by fitting a parabolic function to the lower	(Venek et al. 2017)	Control : 0.36 (n= 504)

	frequencies in the glottal flow spectrum.	(Venek et al. 2014)	Suicide : 0.50 (0.09) (n=84) $p \leq 0.05$ (Venek et al. 2014)
MDQ	The glottal closure instants (GCI) and the dispersion of peaks in relation to the GCI position is averaged across different frequency bands and then normalized to the local glottal period which outputs the MDQ parameter .	(Venek et al. 2014)	N/A

3.2.2 Prosodic features analysis

Among prosodic features F0 is the most used one. France et al. (France et al. 2000) studied the prosodic feature (f0), along with other acoustic features extracted from two recording samples, male subjects sample and female subjects samples. They found that formant and spectral features are the best discriminators between male and female samples, and that prosodic feature is the most discriminator within male subjects. Belouali et al. (Belouali et al. 2021) analysed prosodic features to study the statistical significance of the correlation of these features in army veterans' daily conversation and suicidality. A fused set of 15 acoustic features extracted from the speech were measured by the ensemble feature selection. Random Forest (RF) classifier was applied to determine suicidal group, which yielded 86% sensitivity and 70% specificity. Gideon et al. (Gideon et al. 2019) extracted eGeMAPS features that include prosodic features along with emotional features obtained from the speech from recordings of natural phone conversations to detect suicidal callers. They could extract low-level descriptors (LLDs) for frequency, amplitude, spectral and energy, the parameters in every speech section, which yielded 23 values per frame.

3.3. Formant features

Formant features are also known as filter features, which measure the resonant characteristic of the nasal and vocal tracts that filter the source coming from the vocal folds. The shape of the nasal and vocal tracks are used to reduce some frequencies and increase other frequencies.

3.3.1. Formant features and suicidal

As formant features include vocal tract and acoustic resonance information, hence these features can capture the changes in the suicidal patient's speech changes, if these changes are reflected in mucus secretion and muscle tension. Early studies reported that suicidal speech can be associated with increase in formant frequency and decrease in its bandwidth. In addition to that, suicidal and depressed patients have different, as suicidal patients had shifts in power from lower to higher frequency compared to depressed patients (France et al. 2000). Table 5 lists the most used formant features in automatic suicide assessment.

Table 5 Formant feature used in automatic suicide assessment.

Feature	Description	Study	Significant test
F1	The first peak in the spectrum (especially of voiced utterances such as vowels) results from a resonance of the human vocal tract.	(Venek et al. 2014)(Gideon et al. 2019)(Stasak et al. 2021) (Shah et al. 2019)	Control : 0.1893 (n= 504) Suicide : 0.1805 (n=84) $p \leq 0.001$ (Venek et al. 2014)

		(France et al. 2000)	
F2	The second peak in the spectrum (especially of voiced utterances such as vowels) results from a resonance of the human vocal tract.	(Gideon et al. 2019) (Shah et al. 2019) (France et al. 2000)	N/A
F1 variability	Measures of dispersion of F1 (variance, standard deviation).	(Venek et al. 2017) (Venek et al. 2014)(Gideon et al. 2019) (Stasak et al. 2021) (Shah et al. 2019) (France et al. 2000)	N/A
F2 variability	Measures of dispersion of F2 (variance, standard deviation).	(Venek et al. 2017)(Gideon et al. 2019) (Shah et al. 2019) (France et al. 2000)	N/A
Line spectral features	Measure the linear correlation between spectral frequencies	(Chakravarthula et al. 2020)	N/A
Spectral flex	Measures the flex ratio in spectral.	(Gideon et al. 2019) (Shah et al. 2019)	N/A
Spectral stationarity	Measures the range of the prosodic inventory used over utterances and the monotonicity of the speech	(Scherer et al. 2013)	N/A

3.3.2. Formant features analysis

Many studies have found associations between formant feature changes and suicidal state. For instance, the early study by France et al. (France et al. 2000) showed that the increase in formant features, specifically the formant frequencies and the decrease in formant bandwidth are higher in suicidal patients. Stasak et al (Stasak et al. 2021) extracted the formant F1 features, in addition to the disfluency and voice quality features from the speech of 226 psychiatric inpatients with a recorded suicidal behavior. They manually annotated the recorded speeches and converted them to low-dimensional vectors, which helped to identify suicidal patients with more than 73% accuracy. Shah et al. (Shah et al. 2019) extracted the F0, F1 and F2, spectral flux, loudness and average temporal interval of voiced, as well as silent speech sections. Every feature was

averaged over the total duration of the video yielding scalar features. Variance in pitch was measured as the standard deviation in the F0 values in voiced sections.

3.4. Spectral features

Spectral features measure the spectrum of the subject's speech; which represents the frequency distribution of the speech signal at a given time. Some of the widely used spectral features for suicidal speech analysis include Mel Frequency Cepstral Features (MFCCs) and Power Spectral Density (PSD). Many previous studies have proven the relationship between spectral features, such as energy shift that is measured by PSD, and suicidal speech (Yingthawornsuk et al. 2007). Akkaralaertsest et al. (Akkaralaertsest and Yingthawornsuk 2015) extracted and fused spectral features, including the Glottal Spectral Slope (GSS) and MFCC from the voiced section of the speech sample database. Following that, they combined these spectral features with other acoustic features, mainly source and formant features to a classification model. Their results prove that by combining features from the speech production system (source features), along with spectral features can largely increase the classification accuracy. Keskinpala et al (Keskinpala et al. 2007) extracted and analyzed the spectral features of male and female speech samples from high-risk suicidal patients, specifically, they analyzed MFCC and energy in frequency bands. Their results suggest that mel-cepstral coefficients and energy in frequency band features are highly effective to distinguish between depressed and suicidal patients. Anunvrapong et al. (Anunvrapong and Yingthawornsuk 2014) leveraged MFCC and Delta-MFCC (ΔMFCC) spectral features. They extracted and analyzed ΔMFCCs which is the sixteen consecutive MFCCs, then trained a classifier using various speech datasets of depressed and suicidal patients. Their proposed classifier can distinguish between suicidal and depression patients with 95% accuracy. Ozdas et al. (Ozdas et al. 2004b) extracted MFCC features by dividing the voiced speech segments and measuring the logarithm of the discrete Fourier transform (DFT) of every speech segment. Every log spectrum is further filtered with 16 triangular filters. Using only the MFCC features, they could reach 80% accuracy in classifying near-term suicidal patients and non-depressed subjects. Similarly, Wahidah et al. (NH et al. 2015b) used PSD and MFCC features and achieved 79% accuracy in classifying suicidal patients. Table 6 lists the most used spectral features in automatic suicide assessment.

Table 6 Spectral feature used in automatic suicide assessment.

Feature	Description	Study	Significance test
MFCC	The coefficients derived by computing a spectrum of the log-magnitude Mel-spectrum of the audio segment. The lower coefficients represent the vocal tract filter and the higher coefficients represent periodic vocal fold sources.	(Chakravarthula et al. 2020)(Belouali et al. 2021) (Gideon et al. 2019) (Akkaralaertsest and Yingthawornsuk 2015) (Ozdas et al. 2004b) (Keskinpala et al. 2007) (NH et al. 2015b) (NH et al. 2015a) (Kaymaz Keskinpala et al. 2007)(Reddy et al. 2021)	N/A
ΔMFCC	Measures rapid temporal information captured in the MFCC extraction	(Anunvrapong and	Control: 0.1893 (n= 504) Suicide: 0.1805 (n=84)

		Yingthawornsuk 2014)(Belouali et al. 2021)	$p \leq 0.001$ (Belouali et al. 2021)
PSD	describes how the power of your voice is distributed over frequency	(Yingthawornsuk et al. 2006) (Keskinpala et al. 2007) (NH et al. 2015b) (NH et al. 2015a) (Yingthawornsuk et al. 2007)	N/A
GSS	Measures the periodogram of the voiced segments.	(Akkaralaertsest and Yingthawornsuk 2015)	N/A
PS	The feature is essentially an effective correlate of the spectral slope of the speech signal.	[15] (Venek et al. 2014)	Control : -0.20 (n= 504) Suicide : -0.24 (n=84) $p \leq 0.05$ (Venek et al. 2014)

4. Suicidal visual markers

Visual markers are classified as facial features; eye movement features or posture features.

4.1. Facial features

A growing literature in mental disorder detection has proven that facial appearance and facial expression can carry significant non-verbal cue that can be further interpreted to assess various mental disorders such as depression (Pampouchidou et al. 2017), bipolar (Venn et al. 2004) and social anxiety (Silvia et al. 2006). Given that mental disorders are driving factors of suicide, the facial expression is an important non-verbal suicidal cue for distinguishing suicidal patients (Wang et al. 2021b). Kleiman et al. (Kleiman and Rule 2013) analyzed the facial feature extracted from photos of 40 people who committed suicide. After performing a t-test, they found that accuracy in discerning whether a subject had committed suicide was significantly higher than chance guessing by participants. Laksana et al. (Laksana et al. 2017) extracted various visual features including frowning, smiling, head movement and eyebrow-raising behaviors. They investigated the occurrence and the frequency of these behaviors. The results suggest that facial expressions such as smiling behaviors are more statistically correlated with suicide than other visual features such as eye movements and head movements. Eigbe et al. (Eigbe et al. 2018) studied the smile dynamics (e.g. genuine vs fake smiles) of three groups: people with suicide ideation, people with depression and healthy control subjects. They found that suicidal subjects had the shortest smiling time duration, and the healthy control subject had the longest smiling time. Moreover, the percentage of genuine smiles in suicidal subjects was the lowest, and the healthy control subjects also had a higher percentage of speaking smiles and the longest laughing duration compared to depressed and suicidal groups.

4.2. Eye movement features

Eye movement features have been proven to have a strong correlation with suicidal intent, especially when the suicidal patients are intimidated by a suicide-related question that they often try to avoid answering. In such circumstances, the patients tend to avoid eye contact with the interviewer, with frequent gazing down behaviors. Eigbe et al. (Eigbe et al. 2018) studied the gaze aversion (e.g. looking down) of three groups: people with suicide ideation, people with depression and healthy control subjects. They found that the

depressed group had significantly less gaze down count than the suicidal group and healthy control group. Additionally, suicide group subjects have higher gazing down, they also found that suicidal patients spent a higher frequency of gazing down than control patients, and their gazes time were longer than other groups as well. Shah et al. (Shah et al. 2019) investigated the eye gaze movement and concluded that the frequent shift in eye gaze aversion and eye gaze represents a vital behavioral indicator, as they reflect the subjects' social avoidance, which is strongly correlated with suicidal ideation.

4.3. Posture features

Although posture features extracted from body movement have been proven to be a strong indicator of depression (Pampouchidou et al. 2017), very few works have studied the relationship between body movement and suicidal behaviors. Galatzer-Levy et al. (Galatzer-Levy et al. 2021) used OpenFace to extract the angle of the head's yaw (horizontal movements) and head's pitch (vertical movement) from a video recording of suicidal patients' interviews. After analyzing the head movement along with vocal and facial features, and their correlation with suicidality, their results indicate that head yaw and head pitch variability have significant negative linear relationships with suicidality and that low levels of head movement is strongly correlated with suicide severity. Laksana et al. (Laksana et al. 2017) studied head movements that are interpreted as anxious expressions, such as fidgeting, and looking around the room, and their relationship with suicide ideation. They observed that suicidal patients often exhibit anxious expressions and their high head velocity is higher than healthy control subjects that remain relatively stable during the interview. Table 7 summarizes the most used visual feature for automatic suicide detection.

Table 7 Visual suicidal markers

Class	Feature	Studies
Facial	Smile Intensity	(Laksana et al. 2017)(Pampouchidou et al. 2017)(Scherer et al. 2014)
	Smile duration	(Pampouchidou et al. 2017)(Scherer et al. 2014)
	Duchenne Smile Percentage	(Laksana et al. 2017)
	Sharpness of Smile Onset/Offset	(Laksana et al. 2017)
	Frowning Behavior	(Laksana et al. 2017)
	Eyebrow Raises	(Laksana et al. 2017)
	Facial expression (Action unit)	(Galatzer-Levy et al. 2021)(Williamson et al. 2016)
	Smile	(Eigbe et al. 2018)
	frequent changes in facial expression	(Shah et al. 2019)
	Mouth nose distance	(Wang et al. 2018)
	Mouth landmark/ mouth centroid	(Gupta et al. 2014)
	Facial landmark velocity/acceleration	(Nasir et al. 2016)
	Polynomial fitting	(Nasir et al. 2016)
Eyes	shift in eye gaze	(Shah et al. 2019)
	eye gaze aversion	(Shah et al. 2019)(Alghowinem et al. 2016)
	gazing down behavior	(Eigbe et al. 2018)
	pupil location	(Shah et al. 2019)

	gaze angle	(Shah et al. 2019)
	frequently shifting gaze	(Shah et al. 2019)
	eye pupil movement	(Wang et al. 2018)
	blinking frequency	(Wang et al. 2018)
	Eyebrows and mouth corners movement	(Wang et al. 2018)
	Distance between eyebrows	(Wang et al. 2018)
	Distance upper-lower eyelid	(Gupta et al. 2014)
	Average vertical gaze	(Pampouchidou et al. 2017) (Scherer et al. 2014)
	Vertical eyelid movement	(Alghowinem et al. 2016)
Posture	Head Motion Velocity	(Laksana et al. 2017) (Alghowinem et al. 2016)
	head's pitch (up and down movement)	(Galatzer-Levy et al. 2021) (Alghowinem et al. 2016)
	Head yaw (side to side movement)	(Galatzer-Levy et al. 2021) (Alghowinem et al. 2016)
	STIP features (upper body movement)	(Joshi et al. 2013)(Joshi et al. 2012)

In Table 8, we summarize the above-reviewed research in terms of data source, used acoustic and visual features, analysis method, and targeted suicide type. Descriptive statistical methods tried to gain insight into the different audiovisual features and the presence of audiovisual markers in different groups (e.g. suicidal group vs healthy control group). Machine learning methods on the other hand are more focused on the automatic prediction of suicidal patterns in the input features in raw audiovisual data, even in the absence of context (e.g. patient group). Although we cannot rely on statistical analysis alone to draw general conclusions regarding audiovisual features and their relevance to specific suicidal markers, statistical insights have the potential to boost the training process of machine learning models by reducing the training time and increasing predictive performance. For example, this can be accomplished by leveraging the predictions of multiple different models trained on independent datasets from different clinical environments and aggregating them under a single modeling framework.

Table 8 Audio-visual suicide assessment scheme

Ref	**Data source**	**Acoustic features**	**Visual features**	**Analysis method**	**Targeted suicide type**
(Chakravarthula et al. 2020)	Conversation	MFCCs line spectral frequencies jitter shimmer	N/A	SVM	Passive suicide ideation Suicidal action

(Venek et al. 2017)[6](Venek et al. 2014)	Interview	NAQ QOQ PSP MDQ PS Formants (F1, F2)	N/A	SVM AdaBoostM1	Passive suicide ideation
(Belouali et al. 2021)	Interview	MFCC, Energy, Amplitude, F0, jitter, shimmer, amplitude perturbation quotient, pitch perturbation quotient, logarithmic energy, QOQ and NAQ	N/A	logistic regression (LR); random forest (RF); SVM XGBoost (XGB); k-nearest neighbors (KNN); deep neural network (DNN)	Passive suicide ideation Active suicide ideation Suicidal action
(Nasir et al. 2017)	Interview	pitch energy spectral feature voice quality	N/A	Lyapunov coefficient	Active suicide ideation Suicidal action
(Galatzer-Levy et al. 2021)	Interview	speech prevalence	The angle of the head's pitch (up and down movement) yaw (side to side movement) Facial Action Coding Scheme (FACS)	multiple linear regression	Suicidal action
(Gideon et al. 2019)	Conversation EMA	eGeMAPS features	Emotions	DNN (emotion recognition)	Active suicide ideation Suicidal action
(Stasak et al. 2021)	Reading task	F1 Disfluency GRBASI voice quality	N/A	SVM KNN	Active suicide ideation Suicidal action

(Laksana et al. 2017)	Interview	N/A	Smiling, Frowning Behavior, Eyebrow Raises, Head Motion Velocity	SVM Random Forest, Multinomial Naıve Bayes.	Active suicide ideation Suicidal action
(Shah et al. 2019)	Conversation	F0, F1, F2, loudness, spectral flux, pitch variations	shift in eye gaze eye gaze aversion pupil location gaze angle frequently shifting gaze	Multimodal predictive modelling	Active suicide ideation
(Kleiman and Rule 2013)	Images	N/A	Facial features	Signal detection theory	Suicidal action
(Eigbe et al. 2018)	Interview	N/A	Smile Gazing down behavior	Wilcoxon Rank Sum test	Passive suicide ideation Active suicide ideation Suicidal action
(Scherer et al. 2013)	Interview	Energy f0 Peak slope Spectral stationarity	N/A	HMM SVM	Passive suicide ideation Active suicide ideation Suicidal action
(France et al. 2000)	Interview	f0 Amplitude modulation Formants Power distribution	N/A	Autoregressive model	Passive suicide ideation

(Yingthawornsuk et al. 2006)	Read speech interview	Power spectral densities peak power, peak location, PSD	N/A	Analysis of variance (ANOVA)	Passive suicide ideation
(Ozdas et al. 2004a)	Natural speech	Jitter Glottal flow spectrum	N/A	Maximum Likelihood Classifier	Suicidal action
(Akkaralaertsest and Yingthawornsuk 2015)	Interview Read speech	Glottal Spectral Slope MFCC	N/A	PCA Least Squares (LS)	Passive suicide ideation
(Anunvrapong and Yingthawornsuk 2014)	interview	MFCC	N/A	ML and LMS	Passive suicide ideation
(Yingthawornsuk and Shiavi 2008)	Interview Read speech	Vocal-tract articulation	N/A	GMM	Passive suicide ideation
(Ozdas et al. 2004b)	Interview	MFCC	N/A	GMM	Passive suicide ideation
(Keskinpala et al. 2007)	Interview Read speech	MFCC, Energy in frequency bands (power spectral density).	N/A	Unimodal Gaussian modelling	Passive suicide ideation
(NH et al. 2015b) (NH et al. 2015a)	Interview Read speech	PSD MFCC	N/A	Multiple linear regressions	Passive suicide ideation
(Kaymaz Keskinpala et al. 2007)	Interview Read speech	MFCC	N/A	GMM	Passive suicide ideation
(Yingthawornsuk et al. 2007)	Interview Read speech	PSD Vocal-tract articulation	N/A	GMM	Passive suicide ideation

5. Artificial intelligent suicide detection models

To detect suicidal signs from audiovisual data, most of the existing works applied either statistical analysis to describe the uniqueness of suicidal patients compared to healthy subjects, or machine learning classifiers to infer the commonalities among suicidal patients. Statistical analysis is usually performed to infer meta-knowledge from the extracted audiovisual features. For instance, Doval et al. (Nasir et al. 2017) modeled the speech feature stream as the observed variable of a nonlinear dynamical system, then they calculated the largest Lyapunov coefficient and correlation dimension of the speech series. Gideon et al. (Gideon et al. 2019) investigated the statistical relationship between suicidal ideation and emotion estimated from speech. Specifically, they computed the within-subject standard deviation of each emotion to gauge each emotion's variability. Similarly, Shah et al. (Shah et al. 2019) computed GNorm distance to identify the statistical significance between suicidal and non-suicidal groups. ML classification methods are used to categorize the subjects with similar audiovisual features and are expected to distinguish suicidal patients from healthy subjects. Surprisingly, deep-learning models are rarely used in the literature on suicide detection from audiovisual data, despite their popularity in similar tasks such as depression detection (He and Cao 2018), as deep-learning models have been proven to achieve good results when applied to audiovisual data. Overall, automatic suicide assessment can reduce diagnosis cost, offer instantaneous detection and avoids human biases.

5.1. Linear classifiers

SVM are supervised learning models with associated learning algorithms that analyze data for classification tasks, as well as regression analysis. Although SVM is most used for linear classification tasks, but it can also deal with non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces (Sebald and Bucklew 2000). SVM is by far the most used for classifying suicidal patients (Scherer et al. 2013; Venek et al. 2014, 2017; Laksana et al. 2017; Eigbe et al. 2018; Chakravarthula et al. 2020; Belouali et al. 2021; Stasak et al. 2021)(Reddy et al. 2021), this is due to the fact that SVM requires just small set of training data, which made it desirable for the aforementioned works. Chakravarthula et al. (Chakravarthula et al. 2020) leveraged SVM classifier to categorize the subjects into various classes (risk vs no-risk, degree of risk, non-severe vs severe risk). They applied sample weighting to mitigate class imbalance and tuned hyperparameters such as feature normalization scheme. Scherer et al. [9] compared SVM with HMM to classify interview recorded audio into suicidal/Non-suicidal. They reported that SVM can achieve accuracy of up to 75%, but HMM could achieve up to 81.25% classification accuracy. Similarly, Stasak et al. (Stasak et al. 2021) used an automatic 2-class classification using SVM and non-linear cosine k-nearest neighbor (KNN) algorithm. While some other works (Ozdas et al. 2004b; Yingthawornsuk et al. 2007; Yingthawornsuk and Shiavi 2008) have applied GMM) to model the suicidal patients' class distributions of the extracted audiovisual features.

5.2. Boosting classifiers

Ensemble methods leverage more than one weak learner algorithm to improve generalizability and overcome the drawback of a single weak estimator. Boosting is an ensemble technique primarily used to reduce bias that converts weak learners to strong ones. Boosting classifier gives relatively high accuracy when combined more than weak classifiers such as SVM, for suicide/non suicide groups classification. For example, Venek et al. (Venek et al. 2017)(Venek et al. 2014) proposed a hierarchical classification method with two layers. In the first layer, the suicidal patients and the non-suicidal patients are distinguished using SVM classifier. In the second layer, the suicidal repeaters and the non-repeaters are classified using the ensemble algorithm AdaBoostM1. Using only the patients' features, the classification of suicidal vs. non-suicidal patients achieve an accuracy of 85%. However, by adding the clinician's features the accuracy was improved to 86.7%.

5.3. Decision trees and random forests

Decision trees are non-parametric supervised learning methods that can be used for classification. The rationale behind using tree structure is to create a model that predicts the value of a target variable by learning simple decision rules deduced from the data features. A tree can be viewed as a piecewise constant approximation. A random forest is an ensemble method that combines many decision trees at training time. In the context of classification, the output of the random forest is the class selected by most trees. Belouali et al. (Belouali et al. 2021) applied RF classifier on a fused set of 15 acoustic features extracted from the speech to determine the suicidal group, which yielded 86% sensitivity, 70% specificity. However, Laksana et al. (Laksana et al. 2017) reported that SVM performed better than RF when applied to a single visual feature (smiling).

5.4. Deep neural networks

Deep neural networks refer to a class of supervised/unsupervised algorithms that are modeled to imitate the human brain, that are designed to recognize patterns. In the context of a classification task, DNN reads the data through the input layer of multiple perceptrons, passes it through stacked hidden layers, and finally the output layer determines the class of the input. DNN performs better than conventional classifiers when the dataset is relatively large. DNN have been reported to achieve high performance when dealing with audiovisual data. Belouali et al. (Belouali et al. 2021) compared six different supervised classification algorithms on audiovisual to classify suicidal/non suicidal patients, namely DNN, logistic regression, RF, SVM, XGB and KNN. When applied to acoustic features only, DNN yielded the highest specificity 70% compared to other classifiers, but the lowest specificity 50% when applied to linguistic features only, which confirms that DNN are more suitable for audiovisual data.

In Table 9, we compare the performance of different ML classifiers in the context of suicidal assessment using audiovisual.

Table 9 ML classifiers for suicidal classification

Classifier	Advantage	Drawbacks	Best reported accuracy
SVM	Effective in high dimensional spaces. Can tolerate the lack of data, and even if where the number of features is greater than the number of samples. Memory efficient.	Do not take previous observations into account and are trained on the median and standard deviations of the features over the single utterances. [9] Do not directly provide probability estimates. Prone to overfitting	75% accuracy [9] 60.32 recall (Chakravarthula et al. 2020) 86.7% accuracy (Venek et al. 2017) (SVM-AdaBoostM1)
HMM	Can take advantage of the sequential and dynamic characteristics of the observations and classify each segment on the full frequency sampled feature vector. [9]	Memory consuming	81.25% accuracy [9]
AdaBoostM1	Mitigate the bias of single weak learner Resilient to over-fitting	Sensitive to outliers Difficult to scale up, because every estimator bases its decisions on the previous predictors.	86.7% accuracy (Venek et al. 2017) (SVM-AdaBoostM1)
RF	Simple to understand and to interpret	Prone to overfitting	86% sensitivity (Belouali et al. 2021)

	Can tolerate the lack of data The model can be validated using statistical tests.	Create biased trees if some classes dominate	
DNN	Suitable for audiovisual data	Requires large dataset	70% specificity(Belouali et al. 2021)

6. Discussion, limitations and future directions

In this section, we summarize the finding of this review, and discuss the limitation, as well as potential future directions.

6.1. Discussion

Recording settings: The audiovisual data recording is conducted in a controlled, semi-controlled, or free environment. In controlled settings, the recording is conducted inside an acoustically ideal environment such as an anechoic chamber with a high-quality microphone, and in some cases, the patients were explicitly instructed not to drink caffeinated drinks or alcoholic beverages for several hours before the recording to ensure the use of these did not influence their voice quality. While in semi-controlled environment, the requirement of the soundproof chamber and high-quality microphone is relaxed, therefore the recording can be done remotely, however, the subjects are still instructed to follow some guidelines, such as being in an empty room, or facing a laptop during the recording. Finally, in the free environment, the recording is conducted in any circumstances, usually using smartphones, which gives the advantage of interacting with the patients in natural environments, hence minimizing recall bias and maximizing ecological validity, and allowing the study of micro-processes that influence behavior in real-world contexts (Gratch et al. 2021). On the other hand, this poses a great challenge in processing the voice in an uncontrolled environment. Which involves extra work on noise removal and acoustic feature pre-processing, due to the low quality of smartphone microphones compared to dedicated microphones (Barsties and De Bodt 2015).

Important audiovisual features: Among all the acoustic features, MFCC and F0 is the most used feature that has been strongly correlated with the features and has consistently been observed to change with a speaker's mental state. And when MFCCs are inputted to GMM as an effective mechanism of speech analysis, this combination has been proven to be ideal for classifying either high-low levels of suicidality. Facial expression is the most effective feature that has been correlated with suicide. Most of the facial features used in the reviewed works are dynamic features, where the changes in the facial feature are observed over a period of time, such as smile intensity or eye movement, which require video recording. Therefore, we clearly observe that works that used video have generally high accuracy compared to works that use only static visual features extracted from images. Furthermore, works that used multimodal features from different categories or even within the same category yield better accuracy compared with single modality schemes, e.g., combining multiple audio features from different categories or visual cues from face and upper body for instance.

Standardized audiovisual processing framework: Preprocessing, extracting and manipulation of audiovisual features can pose many challenging tasks. However, the usage of collaborative and freely available repositories for speech and video processing frameworks can facilitate boost research in this field. For acoustic feature processing, the collaborative voice analysis repository for speech technologies (COVAREP) (Degottex et al. 2014) and The Geneva minimalistic acoustic parameter set (GeMAPS) (Eyben et al. 2015) are the most popular tools of acoustic feature manipulation. For the visual feature, one of the prominent tools is OpenFace library, which is capable of facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation.

In Table 10, we summarize the strengths and weaknesses of automatic assessment compared to clinical assessment.

Table 10 Advantages and disadvantages of clinical vs automatic suicide assessment

Method	Advantage	Disadvantage
Clinician assessments	Clinician experience Clinicians can ask for further assessments Questionnaire items are interpretable Clinicians can offer treatment pathway	Relatively expensive (Costs of clinic and clinician) Questionnaires often use ordinal and vague variables (eg, never, sometimes) Time-consuming Prone to clinician's biases (e.g. expertise, culture and race) Difficult to measure and quantify complex features
Automatic assessment	Instantaneous detection Low cost Can be performed remotely and continuously. Avoids human biases and single rater. Can capture multimodal features (e.g. audio-visual) Can process complex features using linear/nonlinear multivariate models, and discover new structure in data. Allows scalability because models can be fast and automated	Requires large datasets for training and testing. Most models have not been validated through clinical trials so far. Models accuracy can be affected by biases in data (eg, race, gender, age, noise)

6.2. Ethics and privacy

Developing AI-based suicide detection models that can be applied to real-time systems, such as video streaming and calling apps, raises serious privacy concerns from a research perspective as well as a deployment perspective. From a research point of view, the users' privacy might be jeopardized throughout the data processing stages. When dealing with publically available datasets, the problem arises when users' personal attributes can be predicted, and the identity of subjects can be revealed. In this context, many jurisdictions require certain conditions for research that can compromise users' privacy. The most common procedure is that the researchers must acquire an ethical approval or exemption from their Institutional Review Board (IRB) before the study. In addition to that, they must obtain informed consent from the participant when possible, protect, and anonymize sensitive data during all research stages. Moreover, they need to be careful when linking data across sites is necessary. Finally, when sharing their data, they need to make sure that other researchers also adhere to the same privacy guidelines (Benton et al. 2017). Researchers can rely on public social media datasets for AI-based suicidal behavior research as long as they ensure the preservation of the confidentiality of users. The privacy concern is more challenging when applying such AI-based solutions in real-time and large-scale scenarios.

6.3. Limitation

The limitations of reviewed works on audio-visual data analysis for suicidal behavior detection can be summarized as follows:

Large-size dataset and population generalization**:** One of the main limitations of some related works is the relatively small size of the dataset used to train the proposed models, both in terms of the number of patients and recorded duration. E.g. n=20 (Galatzer-Levy et al. 2021), n=90 (Shah et al. 2019) , n=97 (Nasir et al. 2017), n=126 (Laksana et al. 2017), n=124 (Chakravarthula et al. 2020). Morever, many of the previous works have focused their studies on specific high-risk groups, and the dataset that they used to generate the results and draw the conclusion is related only to that specific group. For example, army veterans (Belouali et al. 2021)(Nasir et al. 2017), Adolescent (Venek et al. 2014), LGBT (Skerrett et al. 2015). However, acoustic variability is highly affected by linguistic information and speaker characteristics such as age, gender and cultural background. Therefore, the challenges in finding a speech-based marker for suicide require the study of a large population from different groups to tune and generalize the proposed suicide assessment tool.

Lack of public datasets: Most of the relevant public datasets, including the datasets presented in Tabel 2, are general-purpose datasets that have been collected in the context of detecting abnormal behavior that can be latent symptoms of mental illnesses. The lack of public, or even private, datasets that have been collected specially for suicidal behavior detection, is one of the main limitations in this line of research.

6.4. Future directions

Unlike text-based suicidal assessment that has been widely studied, suicide assessment through audio-visual data analysis is still in early development. There are many potential future research directions in this area:

- Conversational AI for suicide assessment: conversational AI agents can engage in a conversation with a user through written text or voice. Conversational AI is an important technology that has the potential to dominate the area of automatic suicide detection and intervention. As suicidal patients may feel more comfortable expressing their thoughts to AI agents than to a human. The future direction in this area is to train AI chatbots on a large data-set of suicide notes, and use this pre-trained model and customize it to the characteristics of each patient.
- Multimodal deep-learning for suicide detection: Deep-learning models have been proven to achieve good results when applied to audiovisual data. Applying deep-learning to multimodal features from different audio-visual categories or even within the same category usually yield better accuracy compared with single modality schemes. For example, conventional neural networks (CNN) are effective when dealing with images, video or audio, while recurrent neural networks (RNN) models such as LSTM and GRU are effective in handling time-series data. One prominent future direction is applying a hybrid CNN-RNN model that has been pre-trained on video interviews of confirmed suicidal patients, such a pre-trained model can capture the time-series markers of suicidal behaviors.

7. Conclusion

In this paper, we have reviewed the recent works that studied suicide ideation and suicidal behavior detection through audiovisual feature analysis, mainly suicidal voice/speech acoustic features analysis and suicidal visual cues. Audiovisual features have been proven to convey significant non-verbal cues that can be further interpreted to assess various mental disorders, however, the literature of audiovisual based suicide assessment is still relatively limited to a few small-scale experiments. Automatic suicide assessment is a promising research direction that is still in the early stages. There is a lack of large datasets that can be used to train machine learning and deep learning models that have been proven effective in other similar tasks. Building an AI-enabled suicide risk assessment tool that observes the patient’s biomarker and audiovisual cues and learns suicide ideation patterns might be a decisive indicator that could help the clinician to reach a decisive judgment.

References

Airas M (2008) TKK Aparat: An environment for voice inverse filtering and parameterization. Logop Phoniatr Vocology 33:49–64

Akkaralaertsest T, Yingthawornsuk T (2015) Comparative analysis of vocal characteristics in speakers with depression and high-risk suicide. Int J Comput Theory Eng 7:448

Alghowinem S, Goecke R, Wagner M, et al (2016) Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. IEEE Trans Affect Comput 9:478–490

Anunvrapong P, Yingthawornsuk T (2014) Characterization of ΔMFCC in depressed speech sample as assessment of suicidal risk. In: International Conference on Advanced Computational Technologies \& Creative Media (ICACTCM '2014). pp 119–123

Barsties B, De Bodt M (2015) Assessment of voice quality: current state-of-the-art. Auris Nasus Larynx 42:183–188

Belouali A, Gupta S, Sourirajan V, et al (2021) Acoustic and language analysis of speech for suicidal ideation among US veterans. BioData Min 14:1–17

Benton A, Coppersmith G, Dredze M (2017) Ethical research protocols for social media health research. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. pp 94–102

Blanchard M, Farber BA (2020) "It is never okay to talk about suicide": patients' reasons for concealing suicidal ideation in psychotherapy. Psychother Res 30:124–136

Castillo-Sánchez G, Marques G, Dorronzoro E, et al (2020) Suicide Risk Assessment Using Machine Learning and Social Networks: a Scoping Review. J Med Syst 44:205. https://doi.org/10.1007/s10916-020-01669-5

Chakravarthula SN, Nasir M, Tseng S-Y, et al (2020) Automatic prediction of suicidal risk in military couples using multimodal interaction cues from couples conversations. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 6539–6543

Cummins N, Scherer S, Krajewski J, et al (2015) A review of depression and suicide risk assessment using speech analysis. Speech Commun 71:10–49

Degottex G, Kane J, Drugman T, et al (2014) COVAREP—A collaborative voice analysis repository for speech technologies. In: 2014 ieee international conference on acoustics, speech and signal processing (icassp). pp 960–964

Dhelim S, Chen L, Aung N, et al (2022) A hybrid personality-aware recommendation system based on personality traits and types models. J Ambient Intell Humaniz Comput 1–14. https://doi.org/10.1007/s12652-022-04200-5

Dhelim S, Ning H, Aung N (2020) ComPath: User Interest Mining in Heterogeneous Signed Social Networks for Internet of People. IEEE Internet Things J 1–1. https://doi.org/10.1109/JIOT.2020.3037109

Doval B, d'Alessandro C, Henrich N (2006) The spectrum of glottal flow models. Acta Acust united with Acust 92:1026–1046

Drugman T, Bozkurt B, Dutoit T (2012) A comparative study of glottal source estimation techniques. Comput Speech \& Lang 26:20–34

Eigbe N, Baltrusaitis T, Morency L-P, Pestian J (2018) Toward Visual Behavior Markers of Suicidal

Ideation. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, pp 530–534

Eyben F, Scherer KR, Schuller BW, et al (2015) The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans Affect Comput 7:190–202

France DJ, Shiavi RG, Silverman S, et al (2000) Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans Biomed Eng 47:829–837

Galatzer-Levy I, Abbas A, Ries A, et al (2021) Validation of Visual and Auditory Digital Markers of Suicidality in Acutely Suicidal Psychiatric Inpatients: Proof-of-Concept Study. J Med Internet Res 23:e25199

Gideon J, Schatten HT, McInnis MG, Provost EM (2019) Emotion recognition from natural phone conversations in individuals with and without recent suicidal ideation. In: Interspeech

Gratch I, Choo T-H, Galfalvy H, et al (2021) Detecting suicidal thoughts: The power of ecological momentary assessment. Depress Anxiety 38:8–16

Gratch J, Artstein R, Lucas G, et al (2014) The distress analysis interview corpus of human and computer interviews. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp 3123–3128

Gupta R, Malandrakis N, Xiao B, et al (2014) Multimodal prediction of affective dimensions and depression in human-computer interactions. In: Proceedings of the 4th international workshop on audio/visual emotion challenge. pp 33–40

He L, Cao C (2018) Automated depression analysis using convolutional neural networks from speech. J Biomed Inform 83:103–111

He L, Niu M, Tiwari P, et al (2022) Deep learning for depression recognition with audiovisual cues: A review. Inf Fusion 80:56–86

Ji S, Pan S, Li X, et al (2020) Suicidal ideation detection: A review of machine learning methods and applications. IEEE Trans Comput Soc Syst

Joshi J, Dhall A, Goecke R, et al (2012) Neural-net classification for spatio-temporal descriptor based depression analysis. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). pp 2634–2638

Joshi J, Goecke R, Parker G, Breakspear M (2013) Can body expressions contribute to automatic depression analysis? In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013

Kane J, Aylett M, Yanushevskaya I, Gobl C (2014) Phonetic feature extraction for context-sensitive glottal source processing. Speech Commun 59:10–21

Kaymaz Keskinpala H, Yingthawornsuk T, Salomon RM, et al (2007) Distinguishing High Risk Suicidal Subjects among Depressed Subjects Using Mel-Frequency Cepstrum Coefficients and Cross Validation Technique. Disting High Risk Suicidal Subj among Depress Subj Using Mel-Frequency Cepstrum Coefficients Cross Valid Tech 1000–1004

Keskinpala HK, Yingthawornsuk T, Wilkes DM, et al (2007) Screening for high risk suicidal states using mel-cepstral coefficients and energy in frequency bands. In: 2007 15th European Signal Processing Conference. pp 2229–2233

Kleiman S, Rule NO (2013) Detecting Suicidality From Facial Appearance. Soc Psychol Personal Sci

4:453–460. https://doi.org/10.1177/1948550612466115

Laksana E, Baltrusaitis T, Morency L-P, Pestian JP (2017) Investigating Facial Behavior Indicators of Suicidal Ideation. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, pp 770–777

Liu RT, Bettis AH, Burke TA (2020) Characterizing the phenomenology of passive suicidal ideation: a systematic review and meta-analysis of its prevalence, psychiatric comorbidity, correlates, and comparisons with active suicidal ideation. Psychol Med 50:367–383

Moher D, Liberati A, Tetzlaff J, et al (2010) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 8:336–341

Nasir M, Baucom BR, Bryan CJ, et al (2017) Complexity in speech and its relation to emotional bond in therapist-patient interactions during suicide risk assessment interviews. In: Interspeech. pp 3296–3300

Nasir M, Jati A, Shivakumar PG, et al (2016) Multimodal and multiresolution depression detection from speech and facial landmark features. In: Proceedings of the 6th international workshop on audio/visual emotion challenge. pp 43–50

NH NNW, Wilkes MD, Salomon RM (2015a) Investigating the Course of Recovery in High Risk Suicide using Power Spectral Density. Asian J Appl Sci 3:

NH NNW, Wilkes MD, Salomon RM (2015b) Timing Patterns of Speech as Potential Indicators of Near-Term Suicidal Risk. Int J Multidiscip Curr Res 3:

Ozdas A, Shiavi RG, Silverman SE, et al (2004a) Investigation of Vocal Jitter and Glottal Flow Spectrum as Possible Cues for Depression and Near-Term Suicidal Risk. IEEE Trans Biomed Eng 51:1530–1540. https://doi.org/10.1109/TBME.2004.827544

Ozdas A, Shiavi RG, Wilkes DM, et al (2004b) Analysis of vocal tract characteristics for near-term suicidal risk assessment. Methods Inf Med 43:36–38

Pampouchidou A, Simos PG, Marias K, et al (2017) Automatic assessment of depression based on visual cues: A systematic review. IEEE Trans Affect Comput 10:445–470

Pestian J, Santel D, Sorter M, et al (2018) A Machine Learning Approach to Identifying Future Suicide Risk. SSRN Electron J. https://doi.org/10.2139/ssrn.3279211

Quatieri TF, Malyska N (2012) Vocal-source biomarkers for depression: A link to psychomotor activity. In: Thirteenth Annual Conference of the International Speech Communication Association

Reddy PP, Suresh C, Rao VK, et al (2021) Vocal Analysis to Predict Suicide Tendency. In: Proceedings of International Conference on Advances in Computer Engineering and Communication Systems. pp 481–488

Ringeval F, Schuller B, Valstar M, et al (2019) AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In: Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop. pp 3–12

Rohlfing ML, Buckley DP, Piraquive J, et al (2021) Hey Siri: How Effective are Common Voice Recognition Systems at Recognizing Dysphonic Voices? Laryngoscope 131:1599–1607

Scherer S, Pestian J, Morency L-P (2013) Investigating the speech characteristics of suicidal adolescents. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp 709–713

Scherer S, Stratou G, Lucas G, et al (2014) Automatic audiovisual behavior descriptors for psychological disorder analysis. Image Vis Comput 32:648–658

Sebald DJ, Bucklew JA (2000) Support vector machine techniques for nonlinear equalization. IEEE Trans signal Process 48:3217–3226

Shah AP, Vaibhav V, Sharma V, et al (2019) Multimodal Behavioral Markers Exploring Suicidal Intent in Social Media Videos. In: 2019 International Conference on Multimodal Interaction. ACM, New York, NY, USA, pp 409–413

Silverman MM, Berman AL, Sanddal ND, et al (2007) Rebuilding the tower of Babel: a revised nomenclature for the study of suicide and suicidal behaviors. Part 2: Suicide-related ideations, communications, and behaviors. Suicide Life-Threatening Behav 37:264–277

Silverman SE, Silverman MK, others (2006) Methods and apparatus for evaluating near-term suicidal risk using vocal parameters

Silvia PJ, Allan WD, Beauchamp DL, et al (2006) Biased recognition of happy facial expressions in social anxiety. J Soc Clin Psychol 25:585–602

Skerrett DM, Kõlves K, De Leo D (2015) Are LGBT populations at a higher risk for suicidal behaviors in Australia? Research findings and implications. J Homosex 62:883–901

Stasak B, Epps J, Schatten HT, et al (2021) Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt. Speech Commun 132:10–20

Sundberg J, Patel S, Bjorkner E, Scherer KR (2011) Interdependencies among voice source parameters in emotional speech. IEEE Trans Affect Comput 2:162–174

Valstar M, Schuller B, Smith K, et al (2013) Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. pp 3–10

Venek V, Scherer S, Morency L-P, et al (2017) Adolescent Suicidal Risk Assessment in Clinician-Patient Interaction. IEEE Trans Affect Comput 8:204–215. https://doi.org/10.1109/TAFFC.2016.2518665

Venek V, Scherer S, Morency L-P, et al (2014) Adolescent suicidal risk assessment in clinician-patient interaction: A study of verbal and acoustic behaviors. In: 2014 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp 277–282

Venn HR, Gray JM, Montagne B, et al (2004) Perception of facial expressions of emotion in bipolar disorder. Bipolar Disord 6:286–293

Walker J, Murphy P (2007) A review of glottal waveform analysis. Prog nonlinear speech Process 1–21

Wang Q, Yang H, Yu Y (2018) Facial expression video analysis for depression detection in Chinese patients. J Vis Commun Image Represent 57:228–233

Wang W, Ning H, Shi F, et al (2021a) A Survey of Hybrid Human-Artificial Intelligence for Social Computing. IEEE Trans Human-Machine Syst

Wang Y, Guobule N, Li M, Li J (2021b) The correlation of facial emotion recognition in patients with drug-na{\"\i}ve depression and suicide ideation. J Affect Disord 295:250–254

Warner CH, Appenzeller GN, Grieger T, et al (2011) Importance of anonymity to encourage honest reporting in mental health screening after combat deployment. Arch Gen Psychiatry 68:1065–1071

Williamson JR, Godoy E, Cha M, et al (2016) Detecting depression using vocal, facial and semantic communication cues. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. pp 11–18

World Health Organization (2014) Preventing suicide: A global imperative. World Health Organization

World Health Organization (2019) Suicide key facts. https://www.who.int/news-room/fact-sheets/detail/suicide. Accessed 14 Apr 2021

Yingthawornsuk T, Keskinpala HK, France D, et al (2006) Objective estimation of suicidal risk using vocal output characteristics. In: Ninth International Conference on Spoken Language Processing

Yingthawornsuk T, Keskinpala HK, Wilkes DM, et al (2007) Direct acoustic feature using iterative EM algorithm and spectral energy for classifying suicidal speech. In: Eighth Annual Conference of the International Speech Communication Association

Yingthawornsuk T, Shiavi RG (2008) Distinguishing depression and suicidal risk in men using GMM based frequency contents of affective vocal tract response. In: 2008 International Conference on Control, Automation And Systems. pp 901–904

Zhang T, Schoene AM, Ji S, Ananiadou S (2022) Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med 5:1–13