# Speech Emotion Recognition in Mental Health: A Systematic Review of Voice Based Applications

Eric Jordan, Motasem Alrahabi, Julien Desclés, Christophe Lemey, Raphaël Terrisse, Marie-Odile Krebs

Submitted to: JMIR Mental Health
on: March 24, 2025

# *Table of Contents*

# Speech Emotion Recognition in Mental Health: A Systematic Review of Voice Based Applications

Eric Jordan[1]; Motasem Alrahabi[1]; Julien Desclés[2]; Christophe Lemey[3]; Raphaël Terrisse[3]; Marie-Odile Krebs[2]

[1] ObTIC Sorbonne Université Paris FR
[2] Institut de Psychiatrie et Neurosciences de Paris, INSERM, Université Paris Cité Paris FR
[3] IRCI, Centre Hospitalier Régional Universitaire de Brest Brest FR

**Corresponding Author:**
Eric Jordan

ObTIC
Sorbonne Université
4 Place Jussieu
Paris
FR



## *Abstract*

**Background:** The nascent field of Speech Emotion Recognition (SER) encompasses a wide variety of approaches with AI technologies providing improvements in recent years. The links between subjects' emotional states and pathological diagnoses are of particular interest. This study aims to investigate the performance of tools combining these approaches with a view to their use within clinical contexts.

**Objective:** The objective of this review was to determine if and how SER technologies have been applied within clinical contexts.

**Methods:** The review includes studies applied to speech (audio) signal, for a select set of pathologies/disorders and only includes those studies that include an evaluation of diagnosis performance using machine learning performance metrics or statistical correlation measures. PubMed, IEEE Explore, ArXiv and Science Direct databases were queried, as recently as February 2025. The QUADAS framework was used to measure the Risk of Bias.

**Results:** A total of 17 articles were included in the final review. The included papers addressed the suicide risk (3), depression (9), psychotic disorders (3) and autism spectrum disorders (2).

**Conclusions:** SER technologies are mostly used indirectly in mental health research and in a wide variety of manners including different architectures, datasets and pathologies. This diversity makes a direct assessment of the technology challenging. Nonetheless, promising results are obtained in various studies that attempt to diagnose patients based on either indirect or direct results from SER models. These results highlight the potential for this technology to be used within a clinical setting. Future work should focus on how these technologies can be used collaboratively by clinicians. Clinical Trial: Prospero ID 1006669

(JMIR Preprints 24/03/2025:74260)
DOI: https://doi.org/10.2196/preprints.74260

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

# Original Manuscript

# Speech Emotion Recognition in Mental Health: A Systematic Review of Voice Based Applications

## Abstract

**Background:** The nascent field of Speech Emotion Recognition (SER) encompasses a wide variety of approaches with AI technologies providing improvements in recent years. The links between subjects' emotional states and pathological diagnoses are of particular interest. This study aims to investigate the performance of tools combining these approaches with a view to their use within clinical contexts.
**Objective:** The objective of this review was to determine if and how SER technologies have been applied within clinical contexts.
**Methods:** The review includes studies applied to speech (audio) signal, for a select set of pathologies/disorders and only includes those studies that include an evaluation of diagnosis performance using machine learning performance metrics or statistical correlation measures. PubMed, IEEE Explore, ArXiv and Science Direct databases were queried, as recently as February 2025. The QUADAS framework was used to measure the Risk of Bias.
**Results:** A total of 17 articles were included in the final review. The included papers addressed the suicide risk (3), depression (9), psychotic disorders (3) and autism spectrum disorders (2).
**Conclusions:** SER technologies are mostly used indirectly in mental health research and in a wide variety of manners including different architectures, datasets and pathologies. This diversity makes a direct assessment of the technology challenging. Nonetheless, promising results are obtained in various studies that attempt to diagnose patients based on either indirect or direct results from SER models. These results highlight the potential for this technology to be used within a clinical setting. Future work should focus on how these technologies can be used collaboratively by clinicians.
**Trial Registration:** Prospero ID 1006669

**Keywords:** Affective computing; Machine learning; Mental health; Psychology; Psychiatry; Speech emotion recognition; Voice

## Introduction

Emotions play a pivotal role in human interaction and communication, influencing various aspects of social interaction, decision-making, and overall well-being. While emotions can be expressed through multiple modalities, including facial expressions, body language, and gestures, human speech remains one of the most prominent and accessible channels for conveying emotional states.

Intonation, rhythm, pitch, and other acoustic features of speech convey subtle emotional cues, reflecting an individual's psychological well-being [1,2]. In recent years, there has been increasing interest in leveraging advancements in machine learning (ML) to analyze and interpret emotional cues from speech, a field known as speech emotion recognition (SER). SER has garnered significant attention in healthcare settings, particularly in the context of mental health assessment and monitoring. The ability to automatically analyze and interpret emotional cues from speech offers several advantages for improving patient care, early detection of mental health issues, and enhancing the overall healthcare experience [3–6].

This study examines the specific role of emotion recognition in identifying psychiatric disorders. To provide context, we review the advancements, applications, and challenges of SER in healthcare, highlighting its potential in mental health monitoring, suicide prevention, mood, psychotic and autism spectrum disorders. This review emphasizes the advantages of SER, including its non-invasive nature, objective assessment capabilities, and potential for automated analysis, while addressing key challenges such as the need for diverse datasets, interpretability of ML models, and their generalization across populations.

This paper is structured as follows: In the first section, we will introduce the history and advancements in SER. Subsequently, we will review the available SER datasets followed by an overview of existing applications of speech and emotion related technologies in clinical settings. Finally, we will discuss current challenges, limitations and prospects for collaboration between the fields of SER and Mental Health.

## Methods

### PRISMA Review

Recognizing and analyzing emotions is an essential tool for the clinician. The fields of psychiatry and psychology have long recognized that learning and mastering this skill is at the very core of the diagnostic and therapeutic process, not only for the caregiver but also for the patient. However, this clinical skill is neither infallible nor sufficient for systematic precision diagnosis. Several studies have focused on the automated recognition of emotions using a variety of computational techniques. What can these new techniques offer to clinicians in the analysis and interpretation of emotions?

To answer these questions, we conducted a systematic review of literature following PRISMA guidelines (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*). Registered in the PROSPERO system with ID 1006669. The aim of this review is to cover studies adopting an acoustic/speech related analysis of mental health questions, with a particular focus on works involving an analysis of the emotional aspect of speech if and when studies paid attention to the emotional dimension.

## Inclusion and exclusion criteria

The inclusion criteria for this review were as follows:

- Studies containing analysis of speech (i.e. audio) signal.
- Containing either a direct or indirect emotion recognition component, indirect components include analyses that could be interpreted from an emotional point of view, for example the use of the openSMILE (open-source Speech and Music Interpretation by Large-space Extraction) feature sets widely used in SER tasks.
- Said speech was issued from a clinical context.

The exclusion criteria were:

- Audio was only analyzed in conjunction with another modality (e.g. text).
- No diagnosis/prognosis aspect was included, e.g. analyzing emotions in isolation without a correlation to patient conditions or outcomes.
- The pathology being examined could be classified as neurological disorder and not a mental health disorder, e.g. Alzheimer's disease.
- The study itself was a review with no experimental component.

## Search Strategy & Screening Process

Database queries were performed using PubMed, IEEE Explore, ArXiv and Science Direct, up until February 2025, with the following keyword search: ("emotion recognition" OR "affective computing" OR "emotional analysis") AND ("psychiatry" OR "psychology") AND ("speech" OR "voice").

During the screening process, two authors applied the eligibility criteria and selected the studies to be included in the systematic review. In case of doubt, it was established that the article would be submitted to the rest of the review group, and no similar cases were encountered.
An initial screening based on the title/abstract of the search results was performed, removing any articles that did not correspond to the inclusion criteria (no mention of speech/audio, no mention of mental health pathologies or mention mainly of excluded pathologies).

From this initial screening, the remaining articles were studied to determine whether their design gave a direct evaluation of their models' performance in a diagnostic task, measured by either ML metrics (F1 score, accuracy, area under the curve (AUC)) or statistical correlation measures e.g. Spearman correlation). As outlined in the flowchart below the majority of articles excluded in this step did not include an audio component, did not include an evaluation of clinical diagnosis performance of their model or applied their model to patients whose pathology was outside of our scope (sometimes analyzing only healthy control subjects).

After filtering for these criteria, a total of 17 studies were retained. The QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies version 2) framework was used to estimate the risk of bias in these selected studies.

The results section is organized in sub-sections to enhance the reader's understanding, particularly regarding the methodological context of SER. First, we provide a non-exhaustive summary of the technical literature (computer science, software and mathematics) and of the various existing databases. We then present the results of the systematic review.

# Results

## PRISMA Review Results

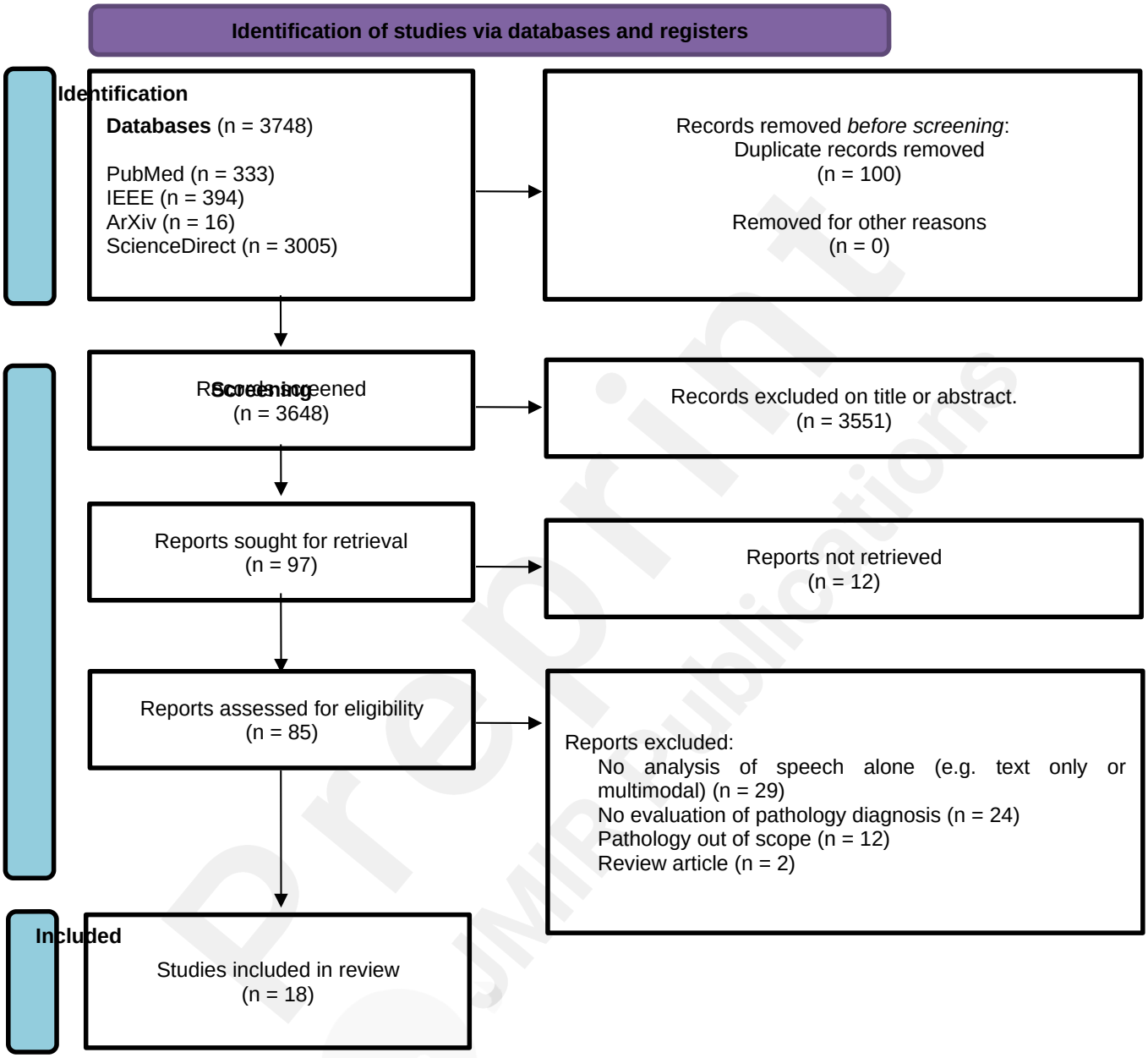



Figure 1. Flow diagram of the PRISMA study selection process

As outlined in Figure 1, a total of 3648 studies were screened with 85 reports retrieved and assessed. From these 85 studies, seventeen were included in the final review with the most common reasons for exclusion being a lack of speech analysis alone (i.e. models using only text or a combination of audio and other modalities) or no diagnostic perspective being assessed within the study (i.e. no prediction of pathological severity or comparison between patients and controls). Of the seventeen studies included in the final selection, 3 addressed suicide risk and suicidal ideation, 9 analyzed depression and mood disorders, 3 studied psychotic disorders and 2 were related to autism spectrum disorders.

Several studies screened did not, in fact, analyze the emotional state of the patients but studied their capacity to detect expressions of emotions in others, i.e. to determine whether their pathologies

impacted their perception of emotional expressions.

### Risk of Bias Assessment

Most of the included studies had a low risk of bias across all fields. As shown in Figure 2, the main exception was in Patient Selection, where 5 studies did not include a control group or did not select from a representative sample [3,7–10]. Additionally, high concerns regarding applicability regarding patient selection were noted in 2 cases [3,8]. 4 studies led to unclear concerns regarding patient selection [5,7,11,12]. The full QUADAS-2 results are available in Multimedia Appendix 1.

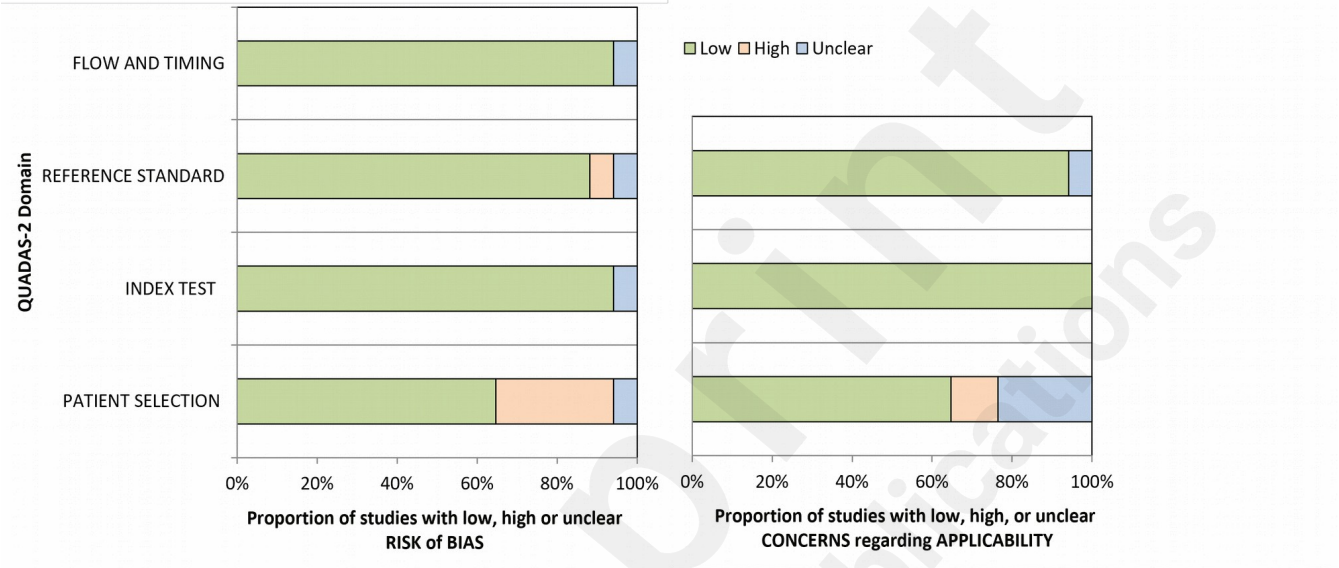



Figure 2. Summary of QUADAS-2 results for selected studies.

This section is divided into two sections. Firstly, we present a narrative review of established methodologies within the field of SER. Secondly, the results of the PRISMA review outlined above are presented.

## Narrative Review

### History of Speech Emotion Recognition

The field of SER has its origins in the 1990s where it emerged as a subsection within the field of speech processing. Initial works approached the task by extracting acoustic features from recordings and then performing statistical analysis using various algorithms to derive correlations between the extracted features and the emotional state of the speaker [13]. The original papers published on the topic suggested that the automatic detection of emotions could be applied in the context of human machine interaction.

These initial publications led to further interest within the field and questions such as the modelling of emotions and the contrast between non-acted and acted emotions. The choice of how to represent emotions is a non-trivial issue and has historically required coordination with the field of psychology to ensure the model used is both compatible with the machines being used for automatic recognition and with the psychological literature on emotions [14].

The models typically used fall into two groups, firstly *categorical models* represent emotions as separate "classes" of emotions (e.g. happy, sad, angry). The "Big Six" model proposed by Ekman et

al.[15] is a well-known example, encompassing happiness, sadness, anger, fear, disgust, and surprise. A more nuanced extension of this categorical approach is Plutchik's model [16], which expands the core emotions by incorporating trust and anticipation, forming an eight-primary-emotion framework. Plutchik's model also introduces the concept of emotional intensity and relationships between emotions, visualized in his 'wheel of emotions' (see Figure 3), where primary emotions blend to form more complex emotional states. Subsequent research has often adapted emotional classifications by adding new categories, removing some, or merging similar emotions into a single class.

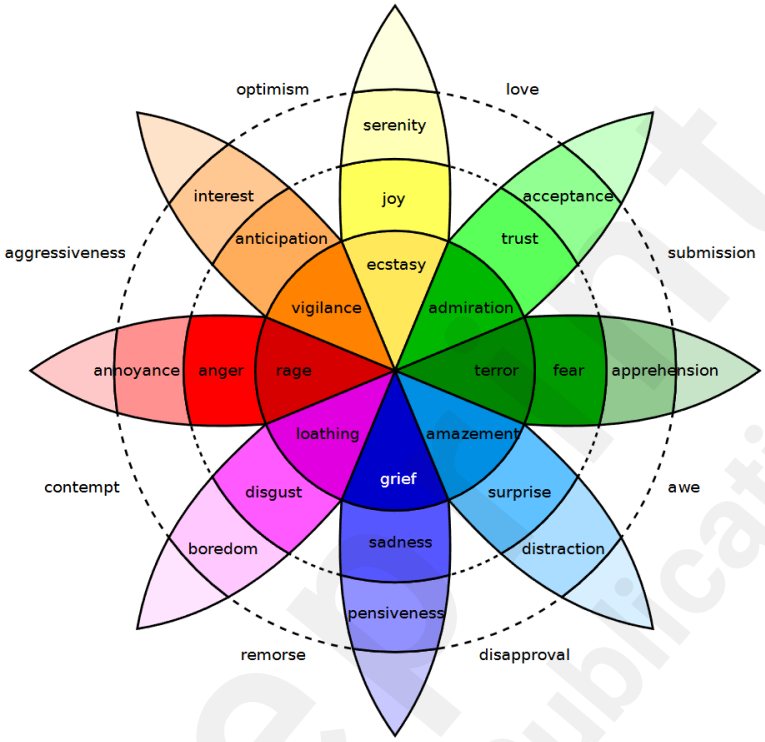



Figure 3. Plutchik's wheel of emotions.

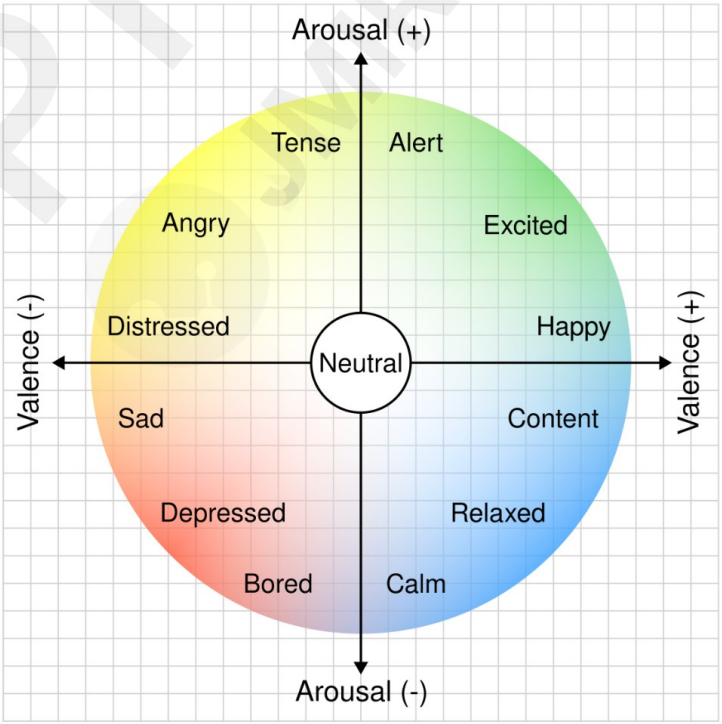



Figure 4. The circumplex model of emotion with its two axes: Valence and Arousal.

Secondly, *dimensional* or *continuous* models interpret emotions as degrees of some underlying features. These models typically involve plotting emotions on two or three axes, the most common of which is the two-dimensional model of arousal and valence [17]. In this model, valence reflects the positive or negative quality of emotion. Happiness has high valence, while sadness has low valence. Arousal reflects the level of physiological activation or intensity associated with emotion. For example, surprise and anger typically have high arousal, while sadness and contentment often have low arousal (see Figure 4).

Much of the existing work in speech emotion recognition predominantly relies on Ekman's 6-scale and Plutchik's 8-scale emotion models, as they offer well-defined categories that facilitate annotation and classification [18]. In contrast, dimensional approaches, which conceptualize emotions along continuous axes, remain comparatively underexplored. However, the dimensional perspective provides a more nuanced framework for capturing emotional complexity and extends beyond speech processing to psychopathology.

Just as emotions can be mapped onto dimensions like valence and arousal, mental health disorders can also be viewed as existing along a continuum rather than within rigid diagnostic categories. This shift is exemplified by the RDoC (Research Domain Criteria) classification in psychiatry, developed by the National Institute of Mental Health, which marks a significant evolution in the approach to mental disorders, by focusing on biological, cognitive and behavioral criteria rather than just clinical symptoms. Its major interest lies in its ability to overcome the limitations of traditional diagnostic classifications, e.g. the DSM (Diagnostic and Statistical Manual of Mental Disorders) and the ICD (International Classification of Diseases), which adopt a categorical approach that is not always relevant to clinical issues. Thus, the RDoC classification proposes a dimensional approach that aims to approach the underlying processes of mental disorders, by tackling transnosographic domains such as emotion, cognition or motivation [19]. This shift from categorical classification to a dimensional approach leads to the study of psychiatric disorders as spectrums of dysfunction, rather than as distinct entities.

## *From Initial Works to Challenges*

By the late 2000s, the field of SER had garnered more widespread interest in the field of speech processing, with a wide variety of feature extraction methods and ML algorithms. However, these developments gave rise to several problems, most notably a lack of comparability between results. Researchers followed a wide range of methods for extracting acoustic features, applying algorithms and as outlined above, there was even a large variation between datasets being used (i.e. emotion modelling and acted versus non-acted emotions). Even some experiments on the same dataset were not comparable due to a lack of standardized training/testing splits.

Facing these challenges in the late 2000s, researchers in the field initiated a series of SER challenges, including the Interspeech 2009 Emotion Challenge [20]. These challenges aimed to standardize key aspects of the SER process, such as datasets, feature extraction methods, and algorithms, to facilitate the generation of comparable results within a unified context. One significant outcome of these efforts was the development of the openSMILE tool for feature extraction [21].

### Arrival of Neural Networks to SER

With the introduction of Convolutional Neural Networks (CNNs) in Automatic Speech Recognition (ASR) in 2012 [22] through their application to spectrograms, these algorithms were quickly adopted for use in Speech Emotion Recognition (SER) contexts [23].

Following trends in both Natural Language Processing (NLP) and ASR, various Neural Network (NN) architectures were also applied. These included end-to-end models, trained on audio from SER datasets and then applied directly to those datasets, supervised pre-trained models, where models that had been trained on labeled data for another task were applied to SER data, finally Self-supervised pre-trained models, using models trained on unrelated, unlabeled data and applying them to SER datasets.

## Established Methodologies

Speech, serving as a primary mode of communication, harbors a wealth of emotional cues that can be harnessed to assess an individual's mental and emotional well-being. Various methodologies have been devised to extract, analyze, and interpret these cues, each providing distinct perspectives on the underlying emotional states.

### Acoustic or prosodic feature extraction

Acoustic or prosodic feature extraction involves the analysis of various acoustic properties of speech signals, such as pitch, intensity, duration, and spectral characteristics [1,2,24,25]. Acoustic features are related to the physical properties of sound waves and can be observed at the level of milliseconds, such as frequency which is related to how humans perceive pitch. Prosodic features are typically observed over a longer timescale, with pitch contours for example measuring how pitch changes over time and other measures including speech rate and pauses. Both sets of features serve as the basis for quantifying emotional cues present in speech and are commonly used as input for ML models. Techniques such as Mel-frequency cepstral coefficients (MFCCs), pitch contour analysis, and formant analysis are commonly employed in acoustic feature extraction for emotion detection tasks [26].

These measures correlate with various aspects of speech production which can be impacted by the emotional state of the speaker. MFCCs, for example, capture the spectral characteristics of speech (i.e. how energy is distributed across different frequencies) that can be linked to certain emotions (higher energy in mid to high frequencies for anger or joy). Pitch contour analysis tracks the variation in the fundamental frequency of speech, that we perceive as pitch, over time. Variation in a speaker's pitch allows us to distinguish between sadness and excitement. Finally, formant values are influenced by the shape and tension of our vocal tract while speaking. Any physiological aspects of emotions (such as smiling) will likely have an impact on these values.

Among the most prevalent tools used for acoustic feature extraction is the *openSMILE* toolkit [21]. This toolkit is available in both .exe form and through Python and contains several widely used collections of acoustic/prosodic measures (referred to as feature sets), allowing for both ease of use and standardization of results.

### Traditional machine learning approaches

Traditional ML approaches involve the use of statistical and pattern recognition algorithms to

classify emotional states based on extracted features. These approaches typically include algorithms such as support vector machines (SVM), k-nearest neighbors (KNN), and decision trees [27–31]. By training these models on labeled datasets of emotional speech samples, they can learn to classify new instances into predefined emotional categories.

### *Recurrent neural networks: LSTM & BiLSTM*

Recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) architectures, gained popularity in speech emotion detection tasks due to their ability to capture temporal dependencies in sequential data [32–35]. These neural networks excel at processing timeseries data such as speech signals, allowing them to capture long-range dependencies and subtle temporal patterns in emotional expression.

### *Transformers*

Transformers, originally developed for NLP tasks, have shown promise in analyzing speech signals for emotion detection [36–40]. These architectures, which include models such as the Bidirectional encoder representations from transformers (known as BERT), excel at capturing contextual information and semantic relationships within sequential data. By fine-tuning pre-trained transformer models on emotion detection tasks, researchers can leverage their powerful language understanding capabilities to analyze emotional content in speech. A study involved the use of Transformer, specifically, a self-attention-based deep learning (DL) model combining a two-dimensional CNN and LSTM network. This model focuses on optimizing feature extraction from speech using MFCCs, achieving an impressive average test accuracy rate of 90% [41].

## Available Datasets

A multitude of datasets for the SER task exist, each falling into one of several categories based on certain characteristics (type of data collected, classification of emotions, relation to any other tasks). However, little to no datasets are available for both SER and Mental Health applications. This section gives an overview of some of the most widely used datasets and their characteristics.

### *DAIC-WOZ*

The DAIC-WOZ dataset is part of the larger Distress Analysis Interview Corpus [42]. It includes audio and transcript recordings of semi-structured clinical interviews between participants and an interviewer. The interviews were designed to detect signs of depression, anxiety, and post-traumatic stress disorder (PTSD). Using the scores from several questionnaires related to psychological distress and current mood the audio and transcript data can be used in a classification of the interviewees.

While the dataset does not explicitly include annotations of the emotional states of the interviewees, labels could be derived using another SER model to directly analyze correlations between the emotional states of the interviewees and their mental states. A total of 193 interviews are available, with each interview lasting between 5 and 20 minutes and all the interviews being conducted in North American English.

### *RAVDESS*

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a multimodal (both audio and face recordings available) database containing both acted speech and song [43]. The

recordings, in North American English, are of 24 professional actors (12 male, 12 female).

Emotions include calm, happy, sad, angry, fearful, surprise and disgust. Each expression (i.e. emotion) produced at two levels of emotional intensity as well as at a neutral expression. The dataset is comprised of a total of 7365 recordings (4320 speech and 3036 singing).

### *FAU-AIBO*

Dataset of spontaneous speech from German children's interaction with a human controlled robot, that they were told was autonomous [44].
The data is annotated on the word level for the presence of 11 emotional states by 5 judges. The emotional labels, chosen based on the emotions observed in the data, are *neutral*, *angry*, *touchy*/*irritated*, *joyful*, *surprised*, *bored*, *helpless*, *motherese* (child addresses robot like a mother would address a child, positive version of reprimanding), *reprimanding*, *emphatic* (speech that is hyper-articulated or accentuated but without any clear accompanying emotion) and *other*.

This dataset was used in the initial 2009 Interspeech SER challenge [20]. Recent work has shown that this dataset presents a significant challenge in achieving strong classification performance even for state of the art methods [23].

### *IEMOCAP*

IEMOCAP is a multi-modal dataset including both audio and motion capture data, allowing for analysis of both speech and facial expressions and gestures [45]. The dataset is comprised of scripted speech as well as improvised scenarios, all produced by actors expressly instructed to produce or elicit certain emotions.

Both categorical and dimensional labelling approaches are used. The categorical labels used are Neutral, Happiness, Sadness, Anger, Surprise, Fear, Disgust, Frustration, Excited and Other. The dimensions used are valence (positive/negative), activation (calm/excited) and dominance (weak/strong). The dataset contains approximately 12 hours of data.

### *CaFE*

CaFE is a Canadian French emotional speech dataset [46]. It includes recordings from six male and six female actors reading six different sentences in Ekman's six basic emotional states (Anger, Disgust, Happiness, Fear, Surprise, Sadness) and in a neutral state, containing approximately 69 minutes of recordings in total. While this dataset may be smaller than other available datasets, it highlights how few datasets exist for languages other than English.

## PRISMA selected studies

The selected studies are presented below, categorized based on mental health pathologies. The studies are grouped in this manner as this is the same approach adopted by the studies that were analyzed, with each study addressing a given pathology. Table S1 (available in Multimedia Appendix 2) presents a full overview of the included studies.

### *Suicide risk and suicidal ideation*

Gerczuk et al. explored differences in speech based on gender in the context of suicide risk, using both interpretable (acoustic) and deep features [47]. The authors reported the best results when using an emotion fine-tuned wav2vec 2.0 model achieving 81% balanced accuracy (i.e. the model had an 81% chance of making a correct prediction for a given example, considering the difference in sample sizes between the different groups) in high-low suicide risk classification. Most notably, this result was achieved by training the model separately on each sex. The authors report a difference in the relationship between acoustic measures and suicide risk between males and females with agitation in males led to increased suicide risk whereas the opposite was true in females, explaining the advantage offered in training the model separately on each group.

Another study analyzed interviews from patients recently released from hospital after suicidal ideation or other circumstances (patients who had attempted suicide, psychiatric patients and healthy controls) [9]. Two separate experiments were conducted, firstly SER classifiers were trained and evaluated using the acoustic features, based on self-reported PANAS-X (Positive and Negative Affect Schedule) emotion labels from the interviews. Secondly, a comparison of the suicidal ideation group and the other groups based on the variability of the reported emotions.
The authors report a maximum AUC of 0.78 when classifying the different emotion labels and an AUC of 0.79 was achieved by using the variability between these emotional states to distinguish between suicidal participants and the other groups. The authors note that the SI group showed lower emotional variability compared to the other groups, indicating that emotional states show promise for use in discrimination between control and pathological groups. These scores demonstrate a good level of discrimination between the groups, with AUC indicating the capacity of a model to differentiate between a randomly selected positive case and a randomly selected negative case (an AUC above 0.7 means the model is achieving a fair discrimination rate, with AUC 0.5 being the equivalent of random chance).

Suicidal ideas among US veterans was investigated by Belouali et al. [3]. This study extracted a large range of features (acoustic, prosodic and linguistic) from recordings of veterans and trained several models (Random Forest (RF), Logistic Regression and Deep Neural Networks (DNNs), among others). Feature selection was applied to use the most relevant features for each model. The best results in classifying suicidal veterans from their non-suicidal counterparts were obtained using a combination of acoustic and linguistic features, achieving a sensitivity of 0.86, specificity of 0.70 and AUC of 0.80. The voices of suicidal individuals were mainly different from non-suicidal counterparts in terms of energy (lower standard deviation of energy contours in voiced segments, lower kurtosis, lower skewness) indicating flatter and less-animated voice. Suicidal individuals were also found to have more monotonous voices.

Overall, studies addressing suicide risk and suicidal ideation achieved good discrimination between the patients and control groups (AUC ~0.8 and accuracy ~80%), indicating that these methods can prove useful in a clinical context to identify patients at high risk of committing suicide.

### Depression and mood disorders

Using both the emotional labels as well as the PHQ-8 depression scores for each participant, Wang et al. attempted to find link between speech features and depression scores as well as depression status (i.e. depressed vs control) [6]. Several models were used: from traditional ML approaches (SVM, RF) to transformer-based approaches (combining several complex models). The best performance was achieved by this complex model, with multiple several models combined, and reaching an accuracy of 77% (i.e. 77% chance of assigning a given participant to the correct group) and F1 score of 0.63 indicates the capability of complex models to outperform more traditional ML approaches on the same dataset.

Yang et al. highlighted how confusion between low mood bipolar depression patients and unipolar depression patients can lead to patients not receiving appropriate treatment [47]. Patients were shown videos to elicit various emotions (happy, sad, disgust, fear, surprise and anger) and asked to respond aloud to several questions. The recordings were labelled (with emotion profiles, i.e. probabilities of each emotion for a given recording) using an SVM classifier trained on the eNTERFACE dataset. Different classifiers were then trained to distinguish between bipolar, unipolar and healthy control groups based on these emotional profiles. The model (combination of both LSTM and BiLSTM architectures) achieved a classification accuracy of 77% when distinguishing between the three groups.

Correlations were found between vocal prosody measures and change in measures of the Hamilton Rating Scale for Depression over the course of 21 weeks as outlined in [48]. Participants were diagnosed with Major Depressive Disorder according to DSM-IV guidelines, and measures of Switching Pause (i.e. time between utterances from patient and interviewer) and fundamental frequency were taken from interviews conducted throughout the duration of the study. When analyzing within subject variation in depression scores the authors found that as depression severity decreased pause duration became shorter and less variable, accounting for 32% of the overall variation over time in patient's depression scores. Furthermore, depression scores obtained from a HLM model were used in linear discriminant classifiers reaching an accuracy score of 69.5% when determining depression severity.

Stepanov et al. address the question of determining depression severity in the Audio-Visual Emotion Challenge (AVEC) 2017 dataset based on speech using low level acoustic features extracted through openSMILE, predicting PHQ-8 scores [49]. The best performance was achieved using these extracted acoustic features to train an LSTM model.

Extracting prosodic features (glottal flow, voice quality and spectral) from the DAIC-WOZ dataset, Mao et al. trained a range of DL models., with the hybrid model achieving an impressive accuracy of 98.7% and an impressive F1 score of 0.987, indicating that the model could almost perfectly distinguish between the control and depression groups [50].

By applying the transformer architecture to frequency related parameters from both the DAIC-WOZ dataset and their own proprietary data, Yang et al [51] achieved maximum F1 scores of 0.78 and 0.87 respectively. By analyzing the frequency components most important to their model’s predictions, the authors found that the frequency range from 600-700 Hz was the most important, corresponding to the Mandarin vowel /e/ or /ê/. The authors present this as a potential biomarker for depression.

Studies applied to depression and mood disorders were the most prevalent among the results of this study (9 of the 17 selected articles). A range of different results were achieved, with newer methods

showing a strong improvement compared to older works (69% accuracy in 2013 with only 2 prosodic features [48], to 98% accuracy in 2023 using a range of prosodic features and deep learning methods [50]). These results show that automated methods can be of use to clinicians for diagnosis purposes (notably when determining severity) and can orient clinicians in adapting their therapeutic approaches.

## Psychotic disorders

Chakraborty et al. used the *emobase* feature set low level descriptors from openSMILE to discriminate between a group Schizophrenia patients and a group of healthy controls based on an interview conducted with a psychologist [52]. Additionally, a Negative Symptom Assessment (NSA-16), score from 1-6 is assigned by the psychologist indicating the severity of the negative symptoms of schizophrenia, i.e. mainly motivational and emotional impairments, prediction was also evaluated. A range of classifiers were trained on the data after conducting feature selection. The best performing classifier for patient vs control was a linear SVM classifier using PCA feature selection, reaching 79.49% accuracy (compared to 66.67% for a classifier predicting the majority class). For NSA-16 prediction, classifiers were trained for each of the different items, with accuracy ranging from 62% to 85%. The best performing classifiers were SVMs, KNN and Decision Trees with a range of different feature selection techniques.

Using a proprietary dataset of Schizophrenia patients with and without Formal Thought Disorder (FTD), first-degree relatives and neurotypical controls and (15 of each, 60 total), Çokal et al. elicited spontaneous speech using an image description task [7]. Pauses were measured in the subjects' speech and classified based on both duration of pause, presence or not of a filler word (*umm*, *ehh*, etc.) and the syntactic context in which the pause occurred.
Patients without FTD produced significantly more unfilled pauses (i.e. pauses without a filler word) than controls in both utterance-initial contexts and before embedded clauses and had more pauses before embedded clauses compared to patients with FTD. When compared to the control and first-degree relative groups, patients with FTD produced longer utterance initial pauses.

The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [53], available through openSMILE, was used by de Boer et al. [54] along with a Random Forest classifier in order to distinguish between patients with Schizophrenia Spectrum Disorders patients and healthy controls. The model achieved a classification accuracy of 86% between healthy controls and patients and an accuracy of 74% in patients between those with negative symptoms and those with positive symptoms. They also indicate that their work is a positive step toward validating language features as biomarkers in the field of psychiatry.

Once again, the results from these studies applied in the context of psychosis showed promise in distinguishing between groups as well as among subgroups of patients with schizophrenia. In a clinical context, these methods can notably be of use for patient screening and early detection where clinicians often struggle to clearly orient their patients. It is worth noting here that no works present in the sample studied involved a direct analysis of emotions in the context of psychotic disorders.

## Autism spectrum disorder

Mencattini et al. [55] applied a statistical learning algorithm to classify samples of speech produced by children with developmental disorders and typically developing children, with a total 102 participants with ages ranging from 6-18 years old. Applies an Emotional Modulation Function rather than applying directly on acoustic features to identify different classes. An AUC of 0.79 was

achieved when determining whether a given child belonged to the typically developing group versus the group with developmental disorders. accuracy of 82% when determining whether a given child was achieved for children with autism disorder.

Saakyan et al. analyzed features from interviews with participants, aged 18-63 years old, diagnosed with autism spectrum disorders (ASD) and neuro-typical participants [56]. By training an XGBoost classification algorithm on the eGeMAPS feature set, the authors achieved an accuracy of 70% and an AUC of 0.78 when distinguishing between the two groups (up to 74% accuracy using a multi-modal approach).

While studies applied directly to ASD were the fewest among the results of our search strategy, the results presented here indicate that SER methods can be used to distinguish ASD individuals from neuro-typical individuals. These methods could then be considered for use in aiding diagnosis. While ASD is typically diagnosed in infancy, these tools can contribute to confirming a diagnosis leading to better health outcomes for children with ASD (especially those without intellectual impairment) and avoiding potentially negative outcomes associated with false positives [57].

## Discussion

Within the fields of psychology and psychiatry, the emotional states of patients can be linked to their condition. This is the case both in a pathological sense, with depressed patients being characterized as having low moods associated with sadness, or in a symptomatic sense where a depressed patient's emotional state can vary depending on their current circumstances. While several works exist at the intersection of these fields and the emerging field of SER, the diversity in methodologies, datasets and results make it difficult to determine what the practical application of these results in a clinical context could look like.

Nonetheless, the results presented in the selected studies are promising, with many of the models discussed above differentiating between pathological and healthy control groups at a significant rate (accuracy ~70-80% and AUC ~0.8). The application of these technologies across a wide variety of domains is also promising, indicating that these technologies could be used for early screening (for example in the diagnosis of psychotic disorders, where clinicians can struggle to guide patients early on).

The use of the *openSMILE* toolkit and its various feature sets (eGeMAPS etc.) throughout several of the selected studies highlights the possibility of integrating SER approaches within a clinical context. Furthermore, the results achieved using these feature sets are promising, especially given their ease of use and interpretation.

To the best of our knowledge, our study is the first systematic review that proposes a synthetic overview of SER in psychiatry. This work could help homogenize future studies and help with results' reproducibility and comparability.

## Challenges and Limitations

While SER holds promise for healthcare applications, several challenges and limitations must be addressed to realize its full potential. One such challenge is the exploration of multimodal emotion recognition, which involves integrating multiple sources of emotional cues, including speech, body

language and facial expressions. While SER primarily focuses on analyzing speech signals, incorporating additional modalities could enhance the accuracy and robustness of emotion recognition systems, particularly in complex healthcare settings, as demonstrated by Saakyan et al. [56].

Additionally, the interpretability of ML models and the ethical considerations surrounding the use of sensitive health data present significant challenges in the development and deployment of SER systems in healthcare settings. Ensuring transparency and accountability in the decision-making process of these algorithms is essential to build trust and confidence among healthcare professionals and patients. As outlined above, some complex models (with added complexity meaning clinical explanation of predictions become increasingly difficult) can achieve results superior to those of traditional ML methods applied to acoustic/prosodic features which are more easily interpreted [6]. If these methods are to be applied within a clinical context, a tradeoff must be found between classification performance and interpretability.

Furthermore, the generalization of SER algorithms across diverse populations and cultural contexts remains a significant hurdle. Variations in speech patterns, dialects, and cultural norms can impact the performance of SER systems, highlighting the need for robust validation and adaptation strategies to ensure the reliability and effectiveness of these systems across different demographics [25].

In addition to these limitations, translating NLP results into routine clinical practice in psychiatry, implying rapid, replicable and scalable analysis, presents specific challenges.
First, possible issues with inaccuracy of vocal recording or transcription can be a hindrance for performing detailed analyses when integrating NLP tools into existing clinical workflows. Indeed, although linguistic analyses are generally considered to be fairly robust in terms of the quality of the transcripts [58], conducting fine analysis such as SER may also require higher quality data which is not always available in a clinical context.

Addressing these challenges requires interdisciplinary collaboration between researchers, clinicians, and technologists to develop innovative solutions that prioritize patient privacy, data security, and ethical considerations while harnessing the potential of SER to enhance mental health assessment and treatment in healthcare settings.

## Future Directions

As has been outlined above, there are many works applying various methodologies, from statistical analysis of acoustic measures to ML and DL methods, in the domains of SER and Mental Health. These techniques have been applied to a range of pathologies, conditions and disorders within the field of mental health. Some works have proposed that acoustic measures of patients' speech be developed as biomarkers of pathology as it is the clinician’s perception of these measures that allow for a diagnosis [54]. This highlights these measures' obvious advantage compared to other methods when collaborating in a clinical context: their ease of interpretation.

Among the various applications similarities can be seen, notably in the reuse of certain methods, especially the use of acoustic measures both in the fields of Mental Health and SER, which can be linked back to the emotional state of the patient. Given that the same methods are employed both for the recognition of pathologies and the recognition of subjects' emotional states, the question arises: Could direct automated analysis of patients' emotional states be useful in the investigation of the aforementioned pathologies?

## Proposed pipelines for audio processing

Here two possible approaches are proposed for detection of pathologies from patients' speech. From our review, the vast majority of works adopt the first approach below, mapping directly from a patient's speech to a possible pathology.

### Speech to Pathology

This pipeline is represented in the vast majority of works found within the literature, where feature extraction or ML/DL methods are applied directly to the signal from recordings of subjects.

### Speech to Emotion, Emotion to Pathology

In this approach, speech from a patient would first be analyzed by some SER system (from one of the various methodologies outlined above), then the emotions recognized by this system could then be analyzed to gain an insight into the patient's pathological status. Some works have adopted a similar approach [9], however this approach to the detection of mental health conditions has not been widely explored.

The pipeline from speech to pathology presents a notable advantage in a clinical context, it offers a clear interpretation of why a given classification was chosen. Furthermore, this approach can be used collaboratively with healthcare professionals to facilitate and complement their work.

One possible difficulty with this approach would be the introduction of error by the SER system. In order to mitigate this error, the choice of the SER system will be of the utmost importance, ensuring that the system's design (for example training data) will allow it to perform properly on a given set of data.

### Shared Challenges

From our review, the variety of methods observed (wide range of audio processing methods and feature sets, datasets of different sizes and compositions, the use of different feature sets) led to difficulties in directly comparing different works, even those that were applied to the same pathologies. This difficulty in comparison is very similar to the difficulty encountered in the early days of the field of SER.

Further insistence on the organization of shared challenges could promote sharing of methods, data and results and hence improve the comparability of work in the field just as was the case in the field of SER. Several challenges have been conducted within this field (or adjacent ones), with the most recent including the AVEC 2019 challenge for depression [59] and the Alzheimer’s Dementia Recognition through Spontaneous Speech challenge for Alzheimer's disease [60].

One hurdle to overcome in the organization of these challenges, especially in a context relating to healthcare, is the sharing of confidential data.

## Dimensional approach to pathology

Dimensional approaches to pathology in mental health can offer a more refined perspective on disorders by focusing on symptom severity and underlying mechanisms rather than rigid diagnostic categories.
To be consistent with the papers we reported, we had no choice but to group in this review mental

health diseases according to disease categories. In contrast to this categorical approach, a dimensional approach could allow for different insights [61]. In practice, this paradigm shift is conducive to the identification of (linguistic) biomarkers [62]. This conceptual change in the way of understanding psychiatric disorders aims to facilitate collaborative work by proposing a common framework through which specialists from different fields can study pathological mechanisms.

Thus, the RDoC classification in psychiatry, described previously, proposes a dimensional understanding of mental disorders, making it easier to grasp individual variability and take into account the different clinical presentations between patients suffering from the same pathology from a categorical point of view.
In the same way, Hierarchical Taxonomy of Psychopathology (HiTOP) proposed by Kotov et al. [63] is a promising model, especially for SER. Instead of replacing which replaces traditional categorical classificationssuch as the DSM (Diagnostic and Statistical Manual of Mental Disorders) and the ICD (International Classification of Diseases), with hierarchical model thatHiTOP hierarchically organizes symptoms into spectra and subfactors. ItHiTOP conceptualizes mental disorders along broad dimensions that reflect underlying emotional and behavioral patterns.
Internalized disorders are characterized by self-directed emotional and psychological symptoms, such as anxiety, depression or phobias. Individuals affected by these disorders tend to internalize their suffering, often isolating their emotions. Externalized disorders, on the other hand, manifest themselves in outwardly directed disruptive behaviors, such as aggression, impulsivity or transgression of social rules. Thought disorders (e.g., schizophrenia) are characterized by confusion, detachment from reality, and unusual experiences. By framing psychopathology as a continuum, HiTOP moves beyond discrete diagnostic categories to capture the overlapping nature of emotional and behavioral dysfunctions.

At the same time, the integration of natural language processing (NLP) and emotion recognition into this approach is opening new perspectives. NLP makes it possible to analyze language data by detecting linguistic cues associated with mental disorders, such as depression or anxiety. These analyses make it possible to extract psychological and emotional dimensions, enriching RDoC criteria with behavioral and cognitive data. The use of linguistic data in the RDoC approach could thus improve diagnostic processes, offering a more detailed and dynamic view of psychiatric disorders, and contribute to more personalized treatments, adapted to the specific profiles of each patient. This convergence of neuroscientific and technological advances could truly transform the way psychiatry approaches mental disorders.

## Conclusions

SER has emerged as a promising tool in mental health care, offering potential for early detection, continuous monitoring, and personalized interventions. This approach is based on ML and AI technologies, and involves the recognition of emotions from recordings of patients in a healthcare context, then analysis of these emotions, with possible applications of other ML methods to either the emotions from the model or the use of the output of the SER model as an input into a classifier to distinguish between populations (control/pathology or other classes fitting the dataset at hand). This is this approach we suggest in the framework of the APAISE project.
However, to be successful, interdisciplinary collaboration between computer scientists, psychologists, linguists, and healthcare professionals will be essential to refine these technologies and ensure their precise, non-biased and ethical implementation.
Beyond that, the next challenge for SER applications in mental health will likely be to integrate SER with other clinical data, as well as biological, genetic, and imaging data, in large-scale multimodal analyses to better characterize and predict psychiatric disorders.

## Acknowledgements

We would like to express our heartfelt gratitude to the Fondation de France and the Fondation de l'Avenir for their invaluable support. This work has also been supported by the French government's "Investissements d'Avenir" program, which is managed by the Agence Nationale de la Recherche (ANR), under the reference PsyCARE ANR-18-RHUS-0014.
Our thanks also go to Jean-Marie Tshimula for contributing to a part of the initial version of this document during his contract on the project.

## Conflicts of Interest

None declared.

## Abbreviations

ASR: Automatic speech recognition
AVEC: Audio-visual emotion challenge
BiLSTM: Bidirectional long short-term memory
CNN: Convolutional neural network
DAIC-WOZ: Distress analysis interview corpus
DL: Deep learning
DSM: Diagnostic and statistical manual of mental disorders
eGeMAPs: extended Geneva minimalistic acoustic parameter set
GRU: Gated recurrent unit
HiTOP: Hierarchical taxonomy of psychopathology
ICD: International classification of diseases
KNN: K-Nearest neighbors
LSTM: Long short-term memory
MFCC: Mel-frequency cepstral coefficients
ML: Machine learning
NIMH: National institute of mental health
NLP: Natural language processing
NN: Neural network
openSMILE: open-source speech and music interpretation by large-space extraction
PANAS: Positive and negative affect schedule
PHQ: Patient health questionnaire
PRISMA: Preferred reporting items for systematic reviews and meta-analyses
PTSD: Post-traumatic stress disorder
QUADAS-2: Quality assessment of diagnostic accuracy studies version 2
RDOC: Research domain criteria
RF: Random forest
RNN: Recurrent neural network
SVM: Support vector machines

## References

1. Latif S, Qadir J, Qayyum A, Usama M, Younis S. Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art. IEEE Rev Biomed Eng 2020;14:342–356.

2. Ozseven T, Arpacioglu M. Comparative Performance Analysis of Metaheuristic Feature

Selection Methods for Speech Emotion Recognition. Meas Sci Rev 2024 Apr 1;24(2):72–82. doi: 10.2478/msr-2024-0010

3. Belouali A, Gupta S, Sourirajan V, Yu J, Allen N, Alaoui A, Dutton MA, Reinhard MJ. Acoustic and language analysis of speech for suicidal ideation among US veterans. BioData Min 2021 Feb 2;14(1):11. doi: 10.1186/s13040-021-00245-y

4. Huang K-Y, Wu C-H, Su M-H, Chou C-H. Mood disorder identification using deep bottleneck features of elicited speech. 2017 Asia-Pac Signal Inf Process Assoc Annu Summit Conf APSIPA ASC Kuala Lumpur: IEEE; 2017. p. 1648–1652. doi: 10.1109/APSIPA.2017.8282296

5. Milling M, Baird A, Bartl-Pokorny KD, Liu S, Alcorn AM, Shen J, Tavassoli T, Ainger E, Pellicano E, Pantic M, Cummins N, Schuller BW. Evaluating the Impact of Voice Activity Detection on Speech Emotion Recognition for Autistic Children. Front Comput Sci 2022 Feb 9;4:837269. doi: 10.3389/fcomp.2022.837269

6. Wang H, Liu Y, Zhen X, Tu X. Depression Speech Recognition With a Three-Dimensional Convolutional Network. Front Hum Neurosci 2021 Sep 30;15:713823. doi: 10.3389/fnhum.2021.713823

7. Çokal D, Zimmerer V, Turkington D, Ferrier N, Varley R, Watson S, Hinzen W. Disturbing the rhythm of thought: Speech pausing patterns in schizophrenia, with and without formal thought disorder. Chialvo DR, editor. PLOS ONE 2019 May 31;14(5):e0217404. doi: 10.1371/journal.pone.0217404

8. Gerczuk M, Amiriparian S, Lutz J, Strube W, Papazova I, Hasan A, Schuller BW. Exploring Gender-Specific Speech Patterns in Automatic Suicide Risk Assessment. Interspeech 2024 ISCA; 2024. p. 1095–1099. doi: 10.21437/Interspeech.2024-1097

9. Gideon J, Schatten HT, McInnis MG, Provost EM. Emotion Recognition from Natural Phone Conversations in Individuals with and without Recent Suicidal Ideation. Interspeech 2019 ISCA; 2019. p. 3282–3286. doi: 10.21437/Interspeech.2019-1830

10. Zhou Y, Han W, Yao X, Xue J, Li Z, Li Y. Developing a machine learning model for detecting depression, anxiety, and apathy in older adults with mild cognitive impairment using speech and facial expressions: A cross-sectional observational study. Int J Nurs Stud 2023 Oct;146:104562. doi: 10.1016/j.ijnurstu.2023.104562

11. Hansen L, Zhang Y, Wolf D, Sechidis K, Ladegaard N, Fusaroli R. A generalizable speech emotion recognition model reveals depression and remission. Acta Psychiatr Scand 2022 Feb;145(2):186–199. doi: 10.1111/acps.13388

12. Cummins N, Amiriparian S, Hagerer G, Batliner A, Steidl S, Schuller BW. An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. Proc 25th ACM Int Conf Multimed Mountain View California USA: ACM; 2017. p. 478–484. doi: 10.1145/3123266.3123371

13. Dellaert F, Polzin T, Waibel AH. Recognizing emotion in speech. Proceeding Fourth Int Conf Spok Lang Process ICSLP 96 1996;3:1970–1973 vol.3.

14. Schuller BW. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing

trends. Commun ACM 2018 Apr 24;61(5):90–99. doi: 10.1145/3129340

15. Ekman P, Sorenson ER, Friesen WV. Pan-Cultural Elements in Facial Displays of Emotion. Science 1969;164(3875):86–88. doi: 10.1126/science.164.3875.86

16. Plutchik R. Emotion: A Psychoevolutionary Synthesis. New York: Harper & Row; 1980.

17. Russell JA. A circumplex model of affect. J Pers Soc Psychol 1980 Dec;39(6):1161–1178. doi: 10.1037/h0077714

18. Plaza-del-Arco FM, Curry A, Curry AC, Hovy D. Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions. Proc 2024 Conf Lang Resour Eval LREC-COLING 2024 Torino, Italy: European Language Resource Association (ELRA); 2024. p. 5696–5710. Available from: https://aclanthology.org/2024.lrec-main.506

19. Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Med 2013 Dec;11(1):126. doi: 10.1186/1741-7015-11-126

20. Schuller B, Steidl S, Batliner A. The INTERSPEECH 2009 emotion challenge. Interspeech 2009 ISCA; 2009. p. 312–315. doi: 10.21437/Interspeech.2009-103

21. Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor. Proc 18th ACM Int Conf Multimed Firenze Italy: ACM; 2010. p. 1459–1462. doi: 10.1145/1873951.1874246

22. Abdel-Hamid O, Mohamed A, Jiang H, Penn G. Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition. 2012 IEEE Int Conf Acoust Speech Signal Process ICASSP Kyoto, Japan: IEEE; 2012. p. 4277–4280. doi: 10.1109/ICASSP.2012.6288864

23. Triantafyllopoulos A, Batliner A, Rampp S, Milling M, Schuller B. INTERSPEECH 2009 Emotion Challenge Revisited: Benchmarking 15 Years of Progress in Speech Emotion Recognition. arXiv; 2024. Available from: http://arxiv.org/abs/2406.06401 [accessed Sep 23, 2024]

24. Zhao S, Yang Y, Cohen I, Zhang L. Speech Emotion Recognition Using Auditory Spectrogram and Cepstral Features. 2021 29th Eur Signal Process Conf EUSIPCO Dublin, Ireland: IEEE; 2021. p. 136–140. doi: 10.23919/EUSIPCO54536.2021.9616144

25. De Lope J, Graña M. An ongoing review of speech emotion recognition. Neurocomputing 2023 Apr;528:1–11. doi: 10.1016/j.neucom.2023.01.002

26. Liu GK. Evaluating Gammatone Frequency Cepstral Coefficients with Neural Networks for Emotion Recognition from Speech. arXiv; 2018. Available from: http://arxiv.org/abs/1806.09010 [accessed Sep 18, 2024]

27. Al Dujaili MJ, Ebrahimi-Moghadam A, Fatlawi A. Speech emotion recognition based on SVM and KNN classifications fusion. Int J Electr Comput Eng IJECE 2021 Apr 1;11(2):1259. doi: 10.11591/ijece.v11i2.pp1259-1264

28. Akinpelu S, Viriri S. OPEN Speech emotion classification using attention based network. Sci Rep.

29. Ismaiel W, Alhalangy A, Mohamed AOY, Musa AIA. Deep Learning, Ensemble and Supervised Machine Learning for Arabic Speech Emotion Recognition. Eng Technol Appl Sci Res 2024 Apr 2;14(2):13757–13764. doi: 10.48084/etasr.7134

30. Jawad M, Ebrahimi-Moghadam A. Automatic speech emotion recognition based on hybrid features with ANN, LDA and K_NN classifiers. Multimed Tools Appl 2023 Apr;82:1–19. doi: 10.1007/s11042-023-15413-x

31. Nath S, Shahi AK, Martin T, Choudhury N, Mandal R. Speech Emotion Recognition Using Machine Learning: A Comparative Analysis. SN Comput Sci Berlin, Heidelberg: Springer-Verlag; 2024 Apr;5(4). doi: 10.1007/s42979-024-02656-0

32. Abdelhamid AA, El-Kenawy E-SM, Alotaibi B, Amer GM, Abdelkader MY, Ibrahim A, Eid MM. Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm. IEEE Access 2022;10:49265–49284. doi: 10.1109/ACCESS.2022.3172954

33. Soltau H, Liao H, Sak H. Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition. arXiv; 2016. Available from: http://arxiv.org/abs/1610.09975 [accessed Sep 18, 2024]

34. Senthilkumar N, Karpakam S, Gayathri Devi M, Balakumaresan R, Dhilipkumar P. Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks. Mater Today Proc 2022;57:2180–2184. doi: 10.1016/j.matpr.2021.12.246

35. Etienne C, Fidanza G, Petrovskii A, Devillers L, Schmauch B. CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation. Workshop Speech Music Mind SMM 2018 2018. p. 21–25. doi: 10.21437/SMM.2018-5

36. Lee S, Han DK, Ko H. Fusion-ConvBERT: Parallel Convolution and BERT Fusion for Speech Emotion Recognition. Sensors MDPI AG; 2020 Nov 23;20(22):6688. doi: 10.3390/s20226688

37. Gong Y, Chung Y-A, Glass J. AST: Audio Spectrogram Transformer. arXiv; 2021. Available from: http://arxiv.org/abs/2104.01778 [accessed Sep 17, 2024]

38. Zhang Z, Wu B, Schuller B. Attention-Augmented End-to-End Multi-Task Learning for Emotion Prediction from Speech. arXiv; 2019. Available from: http://arxiv.org/abs/1903.12424 [accessed Sep 17, 2024]

39. Ullah R, Asif M, Shah WA, Anjam F, Ullah I, Khurshaid T, Wuttisittikulkij L, Shah S, Ali SM, Alibakhshikenari M. Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer. Sensors MDPI AG; 2023 Jul 7;23(13):6212. doi: 10.3390/s23136212

40. Yoon S, Byun S, Jung K. Multimodal Speech Emotion Recognition Using Audio and Text. arXiv; 2018. Available from: http://arxiv.org/abs/1810.04635 [accessed Sep 18, 2024]

41. Singh J, Saheer LB, Faust O. Speech Emotion Recognition Using Attention Model. Int J Environ Res Public Health 2023 Mar 14;20(6):5140. doi: 10.3390/ijerph20065140

42. Gratch J, Artstein R, Lucas G, Stratou G, Scherer S, Nazarian A, Wood R, Boberg J, DeVault D,

Marsella S, Traum D, Rizzo S, Morency L-P. The Distress Analysis Interview Corpus of human and computer interviews. 2014;

43. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. Najbauer J, editor. PLOS ONE 2018 May 16;13(5):e0196391. doi: 10.1371/journal.pone.0196391

44. Steidl S. Automatic classification of emotion related user states in spontaneous children's speech. 2009; Available from: https://api.semanticscholar.org/CorpusID:27397944

45. Busso C, Bulut M, Lee C-C, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Eval 2008 Dec;42(4):335–359. doi: 10.1007/s10579-008-9076-6

46. Gournay P, Lahaie O, Lefebvre R. A canadian french emotional speech dataset. Proc 9th ACM Multimed Syst Conf Amsterdam Netherlands: ACM; 2018. p. 399–402. doi: 10.1145/3204949.3208121

47. Yang T-H, Wu C-H, Huang K-Y, Su M-H. Detection of mood disorder using speech emotion profiles and LSTM. 2016 10th Int Symp Chin Spok Lang Process ISCSLP Tianjin, China: IEEE; 2016. p. 1–5. doi: 10.1109/ISCSLP.2016.7918439

48. Yang Y, Fairbairn C, Cohn JF. Detecting Depression Severity from Vocal Prosody. IEEE Trans Affect Comput 2013 Apr;4(2):142–150. doi: 10.1109/T-AFFC.2012.38

49. Stepanov E, Lathuiliere S, Chowdhury SA, Ghosh A, Vieriu R-L, Sebe N, Riccardi G. Depression Severity Estimation from Multiple Modalities. arXiv; 2017. doi: 10.48550/arXiv.1711.06095

50. Mao K, Zhang W, Wang DB, Li A, Jiao R, Zhu Y, Wu B, Zheng T, Qian L, Lyu W, Ye M, Chen J. Prediction of Depression Severity Based on the Prosodic and Semantic Features with Bidirectional LSTM and Time Distributed CNN. IEEE Trans Affect Comput 2023 Jul 1;14(3):2251–2265. doi: 10.1109/TAFFC.2022.3154332

51. Yang W, Liu J, Cao P, Zhu R, Wang Y, Liu JK, Wang F, Zhang X. Attention guided learnable time-domain filterbanks for speech depression detection. Neural Netw 2023 Aug;165:135–149. doi: 10.1016/j.neunet.2023.05.041

52. Chakraborty D, Yang Z, Tahir Y, Maszczyk T, Dauwels J, Thalmann N, Zheng J, Maniam Y, Amirah N, Tan BL, Lee J. Prediction of Negative Symptoms of Schizophrenia from Emotion Related Low-Level Speech Signals. 2018 IEEE Int Conf Acoust Speech Signal Process ICASSP Calgary, AB: IEEE; 2018. p. 6024–6028. doi: 10.1109/ICASSP.2018.8462102

53. Eyben F, Scherer KR, Schuller BW, Sundberg J, Andre E, Busso C, Devillers LY, Epps J, Laukka P, Narayanan SS, Truong KP. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Trans Affect Comput 2016 Apr 1;7(2):190–202. doi: 10.1109/TAFFC.2015.2457417

54. De Boer JN, Voppel AE, Brederoo SG, Schnack HG, Truong KP, Wijnen FNK, Sommer IEC. Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-

recognition tool. Psychol Med 2023 Mar;53(4):1302–1312. doi: 10.1017/S0033291721002804

55. Mencattini A, Mosciano F, Comes MC, Di Gregorio T, Raguso G, Daprati E, Ringeval F, Schuller B, Di Natale C, Martinelli E. An emotional modulation model as signature for the identification of children developmental disorders. Sci Rep 2018 Sep 27;8(1):14487. doi: 10.1038/s41598-018-32454-7

56. Saakyan W, Norden M, Herrmann L, Kirsch S, Lin M, Guendelman S, Dziobek I, Drimalla H. On Scalable and Interpretable Autism Detection from Social Interaction Behavior. 2023 11th Int Conf Affect Comput Intell Interact ACII Cambridge, MA, USA: IEEE; 2023. p. 1–8. doi: 10.1109/ACII59096.2023.10388157

57. Okoye C, Obialo-Ibeawuchi CM, Obajeun OA, Sarwar S, Tawfik C, Waleed MS, Wasim AU, Mohamoud I, Afolayan AY, Mbaezue RN. Early Diagnosis of Autism Spectrum Disorder: A Review and Analysis of the Risks and Benefits. Cureus 2023 Aug 9; doi: 10.7759/cureus.43226

58. Corcoran CM, Mittal VA, Bearden CE, E. Gur R, Hitczenko K, Bilgrami Z, Savic A, Cecchi GA, Wolff P. Language as a biomarker for psychosis: A natural language processing approach. Schizophr Res 2020 Dec;226:158–166. doi: 10.1016/j.schres.2020.04.032

59. Ringeval F, Schuller B, Valstar M, Cummins Ni, Cowie R, Tavabi L, Schmitt M, Alisamir S, Amiriparian S, Messner E-M, Song S, Liu S, Zhao Z, Mallol-Ragolta A, Ren Z, Soleymani M, Pantic M. AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. arXiv; 2019. doi: 10.48550/arXiv.1907.11510

60. Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. arXiv; 2020. Available from: http://arxiv.org/abs/2004.06833 [accessed Jul 17, 2024]

61. Insel TR, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, Sanislow C, Wang P. Research Domain Criteria (RDoC) : Toward a New Classification Framework for Research on Mental Disorders. Am J Psychiatry 2010;167(7):748–51. doi: 10.1176/appi.ajp.2010.09091379

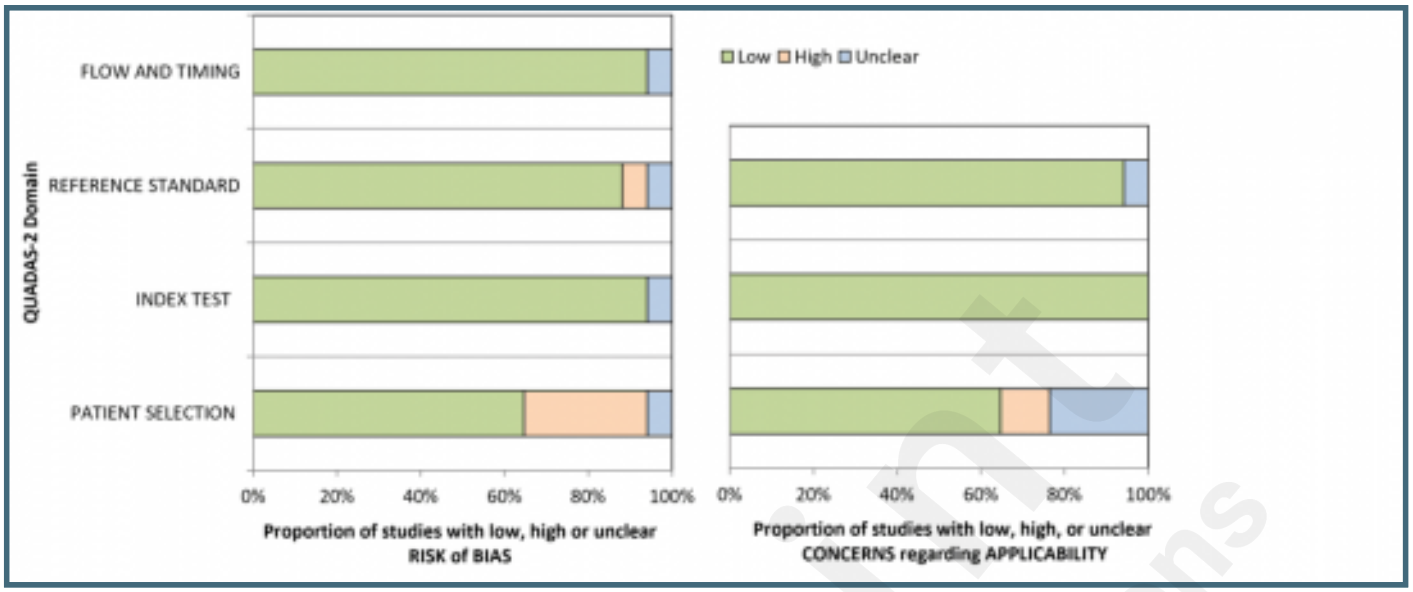
62. Kelly J, Clarke G, Cryan J, Dinan TG. Dimensional thinking in psychiatry in the era of the Research Domain Criteria (RDoC). Ir J Psychol Med 2018 Jun;35(2):89–94. doi: 10.1017/ipm.2017.7

63. Kotov R, Krueger RF, Watson D, Cicero DC, Simms LJ, Markon KE, Wright AGC, Samuel DB, Hopwood CJ, Bagby RM, Jonas K, Morey LC, Bates JH, Patrick CJ, Clark LA, Tackett JR, Tackett JV, Jonas KA. The Hierarchical Taxonomy of Psychopathology (HiTOP): A Dimensional Alternative to Traditional Nosologies. J Abnorm Psychol 2017;126(4):454–477. doi: 10.1037/abn0000258

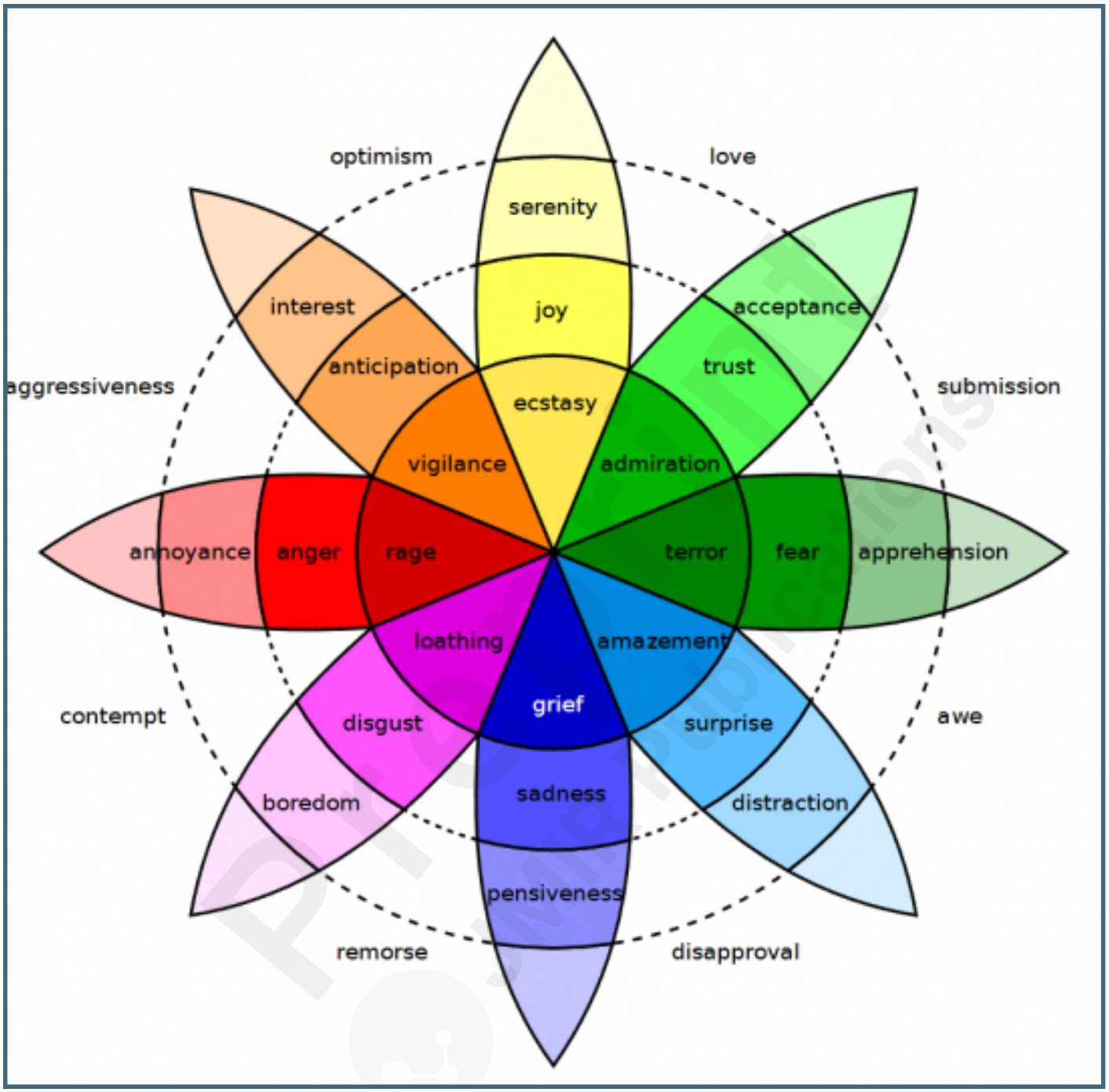
# Supplementary Files

# Figures

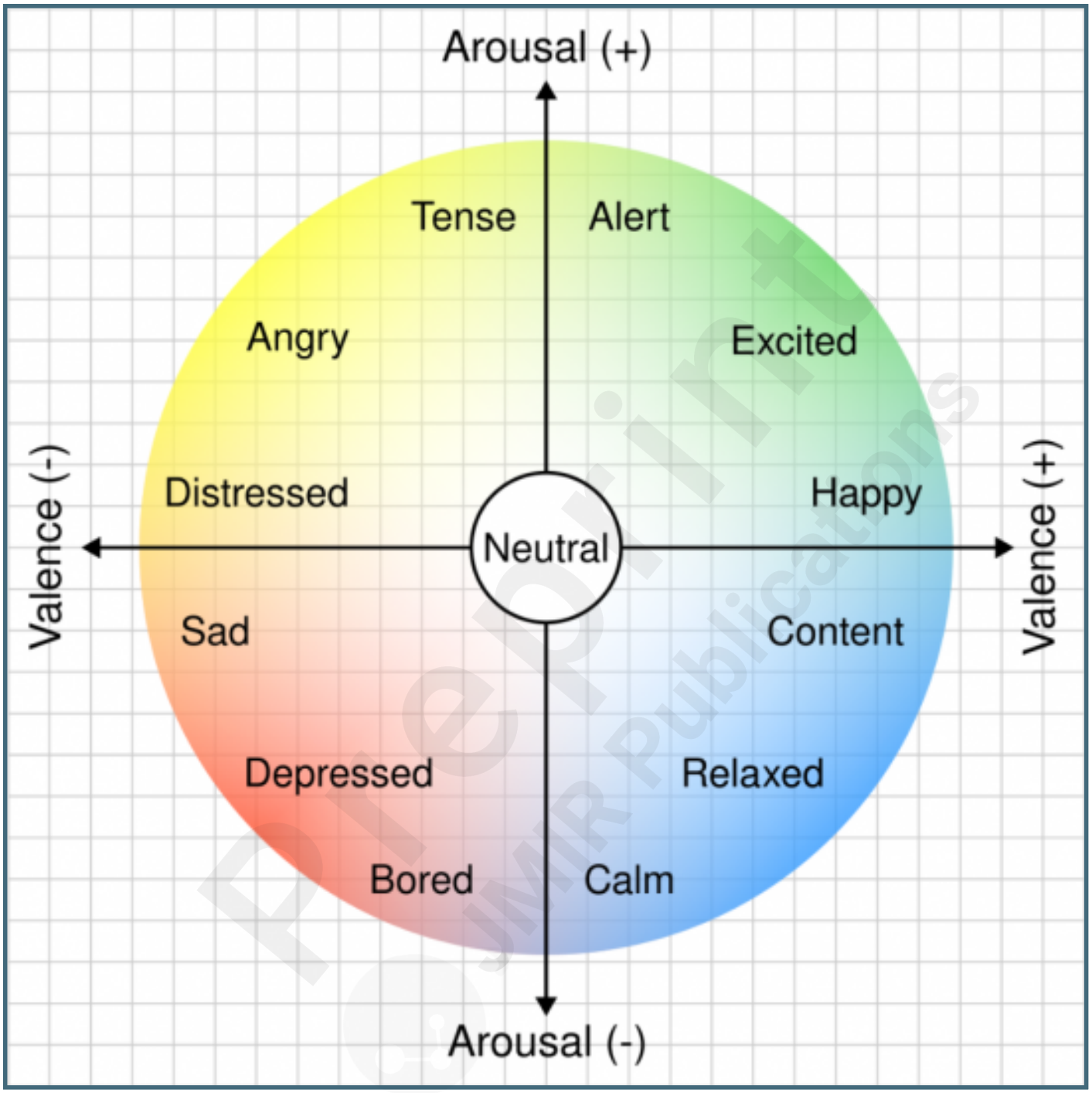
Summary of QUADAS-2 results for selected studies.



Plutchik's wheel of emotions.



The circumplex model of emotion with its two axes Valence and Arousal.



# Multimedia Appendixes

Full QUADAS-2 results.
URL: http://asset.jmir.pub/assets/f2f3aa00ffb4dbf83432dbde1ff90d7d.xlsx

PRISMA selected studies overview table.
URL: http://asset.jmir.pub/assets/36aca1ba71fdd9e015c0ff8328e8cf64.pdf