Spoken Response-Based Automatic Screening of Suicidal Risk Among Adolescents

Rohan Kumar Gupta
Department of Networking and Communications, School of Computing
SRM Institute of Science and Technology
Kattankulathur, Chennai, India
rohankug@srmist.edu.in

***Abstract*—Automatic detection of suicidal risk has become the need of the hour, especially for adolescents. This study presents a suicidal risk classification system. The database utilized in this work is introduced in the 1st SpeechWellness Challenge organized in 2025. It comprises participants' spoken responses to three distinct tasks. The provided baseline systems are based on both conventional and advanced deep-learning models. The implemented deep learning model has provided better detection performance. In this study, we explore three distinct hand-crafted feature sets and their fusion at the decision and feature levels to implement a light-weight suicidal risk classification system. We also proposed a simple method that address the limitation of the baseline conventional model. Moreover, the detection is performed at the segment level, and the impact of different input segment lengths is analyzed. The results show that our proposed method outperformed the baselines on the development set in terms of accuracy. On the test set, our submissions are found to be better than the baselines in terms of recall metric.**

***Index Terms*—suicidal risk detection, conventional machine learning, speech processing**

I. Introduction

In today's competitive world, many individuals prioritize work while often overlooking their health. Among various health concerns, mental health is the most commonly neglected. Emotion regulation refers to an individual's ability to manage and adjust their emotions based on different situations [1]. Difficulty in emotion regulation contributes to the development of mental health conditions, with suicide being the most severe consequence of impaired emotion regulation ability [2]. Clinical interviews and self-report questionnaires are commonly used for diagnostic purposes in clinical practice. During interviews, clinicians evaluate a patient's symptoms through direct interaction to diagnose mental health conditions. However, such interviews demand significant resources, in particular, considerable human effort and the expertise of trained professionals. On the other hand, questionnaires offer a more cost-effective screening alternative, but they are susceptible to limitations such as bias, misrepresentation, and potential respondent deception. These drawbacks can be mitigated by developing an automatic detector for suicidal risk assessment.

One of the challenges of the 2025 Interspeech Conference is the 1st SpeechWellness Challenge, which involves the automatic detection of suicidal risk [3]. The database utilized in this challenge comprises participants' spoken responses to three distinct tasks. The created baselines by the organizers of the challenge involve conventional machine learning and deep learning based binary classifiers. These classifiers are constructed either using data from a single task or by aggregating the predictions of three separate classifiers – each trained on data from one of the three tasks – through the majority voting (MV) approach. The results show that upon combining the three classifiers, the detection performance degrades in most cases compared to the detection performance of individual classifiers. In this study, we proposed an alternative method to address this problem. In addition, we explore three distinct hand-crafted feature sets and their fusion both at the decision and feature level for developing a light-weight suicidal risk detection system. Moreover, the detection is performed at the segment level, and the impact of different input segment lengths is analyzed. The rest of the paper is organized as follows. Section II provided a brief discussion of reported works for suicidal risk detection. The database provided by the organizers of the challenge is described in Section III. Our proposed methodology is presented in Section IV. The experimental results and their brief discussion are provided in Section V. The concluding remarks of this study are included in Section VI.

II. Related works

The audio databases used for developing suicidal risk screening modules have primarily been collected through either an interview setup or by having participants read aloud a predefined text. In the interview setup, the interaction between the participant and the interviewer is recorded. The interview is either structured [4], semi-structured [5] or unstructured [6], [7]. In the literature, several studies employed conventional machine learning-based models. In those studies, different low-level descriptors (LLDs) of the spoken responses have been explored. Those LLDs mainly are spectral [8], prosodic [9], or source [9], [10] features. The extracted spectral features include mel frequency cepstral coefficients (MFCC) [7], [11], [12], formant frequencies [5], [8], power spectral density [11], line spectral frequencies [7]. The analyzed prosodic features consist of fundamental frequency [4], [6], [9], acoustic intensity [7], speech rate [13], pause duration [13]. Additionally, the widely used source features include jitter [7], [9], shimmer [7], [9], normalized amplitude quotient (NAQ) [4], [5], [12], quasi-open quotient

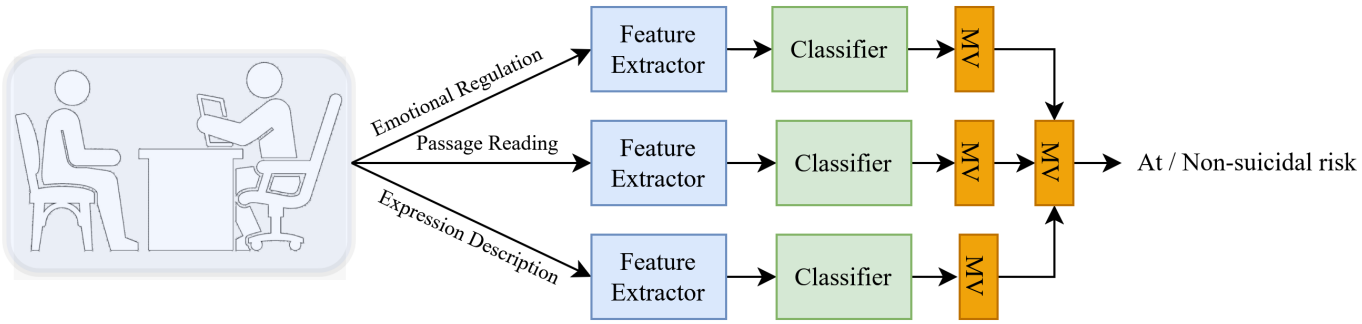

Fig. 1. Schematic diagram of the proposed module for automatic screening of suicidal risk.

(QOQ) [4], [5], [12], slope of glottal flow spectrum [6], [9]. A combination of different types of LLDs is also explored [6], [12]. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), proposed for emotion recognition in [14], has been utilized for suicidal risk detection [6], [15]. It is also utilized for creating one of the baselines for the current challenge [3]. The eGeMAPS comprises a set of LLDs that include spectral, prosodic, and source features. The dimensionality of the set of LLDs is 23. Several statistical functionals are also applied to those LLDs, which make the eGeMAPS of 88-dimensional.

Deep-learning-based feature representations have provided better classification/regression performance in several domains than hand-crafted feature sets. In a recent study [16], the feature representation generated by wav2vec 2.0 (w2v2) model [17] – fine-tuned for estimating dimensional emotions and performed gender-specific modeling– is utilized for assessing suicidal risk detection. The generated feature representation outperformed two hand-crafted feature sets, including eGeMAPS. The authors in [18] have explored the fusion of speech and text embeddings generated using deep-learning models for detecting suicidal risk. The fusion has provided better detection performance than the individual embeddings. This kind of fusion is also explored for creating one of the baselines for the current challenge [3]. In that baseline, the speech and text encoders are w2v2 and Bidirectional Encoder Representations from Transformers (BERT) [19], respectively.

III. Database

This study utilizes the database provided by the challenge organizers [3]. The database comprises spoken responses of 600 participants (180 males and 420 females) to three tasks. The participants are school students in China aged 10-18 years. Those three tasks are referred to as emotion regulation (ER), passage reading (PR), and expression description (ED). In the ER task, the participants responded to a question related to their experience and ability to regulate emotional distress. In the PR task, the participants are asked to read out a text that is emotionally neutral. In the ED task, the participants described a woman's facial expression in the given picture. The collected spoken responses are labeled for assessing the suicidal risk using the Mini International Neuropsychiatric Interview for Children and Adolescents (MINI-KID) suicide diagnostic [20]. The 600 participants in the database are equally distributed in two classes: at suicidal risk or non-suicidal risk.

For the challenge, the organizers divided the database into training, development, and test sets comprising data from 400, 100, and 100 participants, respectively. All three sets are class-balanced. The distribution of spoken responses among three tasks in each set is shown in Table I. It can be observed from the table that data across three tasks are quite different for a set. The average duration in seconds for ER, PR, and ED tasks are 36.4, 47.4, and 23.5, respectively.

TABLE I
DISTRIBUTION OF SPOKEN RESPONSES IN HOURS FOR THE CONSIDERED TASKS IN THREE SETS.

Task	Training	Development	Test
ER	4.10	0.97	1.01
PR	5.25	1.32	1.32
ED	2.65	0.63	0.64

IV. Methodology

Our proposed method addresses the drawback of the created baselines by the challenge organizers [3]. In the baseline models, when combining the three tasks, the predictions from three separate classifiers – each trained on data from one of the three tasks – are concatenated before applying majority voting (MV). As outlined in Section III, the duration of spoken responses varies significantly across all three tasks. We hypothesize that this disparity may introduce a bias, potentially leading to a decline in detection performance compared to the results achieved by individual tasks in most cases. To mitigate this bias, we applied the MV approach twice instead of once. We applied the MV at the decision level of each task as well as on the accumulated decisions of three tasks, as shown in Figure 1.

In this work, to develop the detector on data from an individual task, we analyzed the efficacy of different input segment lengths varying from 2 to 10 seconds in steps of 2 seconds with a shift of 25% of the segment length. Additionally, we utilized three different hand-crafted feature sets: eGeMAPS [14],

MFCC appended with its first- and second-order derivatives, and the feature set utilized in 2016 Audio/Visual Emotion Challenge (AVEC) for depression detection [21]. The resultant dimensionality of the MFCC feature set is 39. We refer to the feature set from AVEC 2016 as AVEC-16 in the remainder of the paper. The AVEC-16 comprises different LLDs, which include spectral, prosodic, and source features, and is 79-dimensional. The LLDs are computed with a frameshift of 10 msec. To calculate the high-level descriptors of the MFCC and AVEC-16 feature sets, we applied the mean function over a chosen segment length. We also combined these feature sets both at the decision and feature levels. For the development of the detector, we utilized a support vector machine (SVM) based classifier from the sklearn library. All hyperparameters of the classifier are set at the default values.

V. Results and Discussions

This section presents the experimental results for detecting suicidal risk on the development and test sets of the challenge database. Detection is carried out on spoken responses from three separate tasks individually, as well as by merging the outputs of the three detectors at the decision level. For each task, three distinct feature sets are explored: eGeMAPS, MFCC, and AVEC-16. To further improve the detection performance, we combined these feature sets at decision and feature levels. All the experiments are performed on Spyder IDE (v3.3.6) with Python 3.7 running on an Intel core i7 8th Gen. processor. The organizers of the challenge chose the accuracy metric to evaluate the efficacy of the detector.

A. Decision level fusion

Table II shows the detection performances of suicidal risk for three considered feature sets at five different input segment lengths. It can be observed from the table that our proposed method has outperformed the best-performing baseline by an absolute difference of 6%. Figure 2 shows the distribution of detection performances for each task and their combination from Table II. It can be observed that the combination of three tasks provides better detection performance than individual tasks in the overall sense. In addition, the impact of segment lengths on the detection performance is minimal for the combined case.

TABLE II
Detection performances of suicidal risk in terms of accuracy on the development set.

Model	Chunk size	ER	PR	ED	Combined
Baseline (eGeMAPS+SVM)		0.52	0.54	**0.58**	0.53
Baseline (W2V2+BERT)		0.56	0.57	0.54	0.56
eGeMAPS+SVM	2 sec	0.54	0.53	0.54	0.53
	4 sec	0.60	0.48	0.59	0.63
	6 sec	0.60	0.51	0.57	0.58
	8 sec	0.63	0.55	0.60	**0.64**
	10 sec	0.61	0.55	0.60	0.61
MFCC+SVM	2 sec	0.61	0.53	0.55	0.59
	4 sec	0.58	0.57	0.60	0.63
	6 sec	0.56	0.50	0.60	0.61
	8 sec	0.59	0.55	0.59	**0.64**
	10 sec	0.57	0.51	0.55	0.60
AVEC-16+SVM	2 sec	0.56	0.55	0.59	0.62
	4 sec	0.57	0.58	0.62	**0.64**
	6 sec	0.57	0.57	0.57	0.61
	8 sec	0.58	0.58	0.58	0.61
	10 sec	0.59	0.58	0.62	0.62

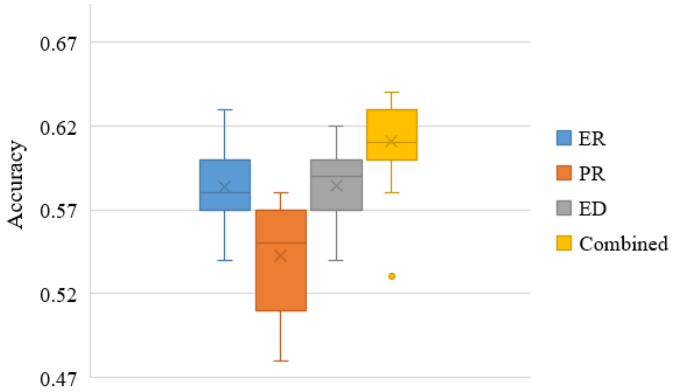

Fig. 2. Distribution of detection performances across all evaluated models for both individual and combined tasks.

TABLE III
Detection performances of suicidal risk upon decision level fusion of considered feature sets.

Feature sets	Accuracy	Precision	Recall
eGeMAPS+MFCC	**0.66**	0.66	0.66
eGeMAPS+AVEC-16	0.63	0.63	0.63
MFCC+AVEC-16	0.64	0.65	0.64
eGeMAPS+MFCC+AVEC-16	0.65	0.65	0.65

To further improve the detection performance, we fused the considered feature sets at the decision level. For the same, we applied the MV on the accumulated predictions across all considered feature sets and all three tasks. For each feature set, the best-performing model in Table II is considered. Table III demonstrates that the fusion of eGeMAPS and MFCC has further improved the detection performance by 2%.

B. Feature level fusion

Table IV shows the detection performances of suicidal risk upon feature level fusion of the considered feature sets. The feature sets are simply concatenated at the segment level. It can be observed from the table that MFCC+AVEC-16 and eGeMAPS+MFCC+AVEC-16 yielded the same detection performance as found for the best-performing model at the decision level fusion (eGeMAPS+MFCC). The overall best detection performance is noted for the fusion of all three feature sets and for data from the ED task only. The absolute increment compared to the best-performing baseline is found to be 9%.

Figure 3 shows the distribution of detection performances for each task and their combination from Table IV. It can be observed that the combination of three tasks provides better detection performance than individual tasks in the overall

TABLE IV
DETECTION PERFORMANCES OF SUICIDAL RISK IN TERMS OF ACCURACY UPON FEATURE LEVEL FUSION OF CONSIDERED FEATURE SETS.

Model	Chunk size	ER	PR	ED	Combined
Baseline (eGeMAPS+SVM)		0.52	0.54	**0.58**	0.53
Baseline (W2V2+BERT)		0.56	0.57	0.54	0.56
eGeMAPS+ MFCC+SVM	2 sec	0.56	0.54	0.53	0.59
	4 sec	0.58	0.49	0.59	0.61
	6 sec	0.61	0.53	0.60	0.62
	8 sec	0.59	0.54	0.58	**0.65**
	10 sec	0.57	0.57	0.61	0.57
eGeMAPS+ AVEC-16+SVM	2 sec	0.61	0.51	0.57	0.63
	4 sec	0.63	0.51	0.58	**0.64**
	6 sec	0.57	0.57	0.53	0.57
	8 sec	0.57	0.55	0.59	0.62
	10 sec	0.58	0.58	0.62	0.63
MFCC+ AVEC-16+SVM	2 sec	0.57	0.56	0.60	0.65
	4 sec	0.58	0.57	0.64	0.65
	6 sec	0.57	0.55	0.58	0.59
	8 sec	0.56	0.57	0.62	0.62
	10 sec	0.62	0.58	0.62	**0.66**
eGeMAPS+ MFCC+ AVEC-16+SVM	2 sec	0.60	0.51	0.57	0.62
	4 sec	0.62	0.52	0.59	0.66
	6 sec	0.62	0.53	0.59	0.63
	8 sec	0.53	0.56	0.61	0.61
	10 sec	0.58	0.59	**0.67**	0.65

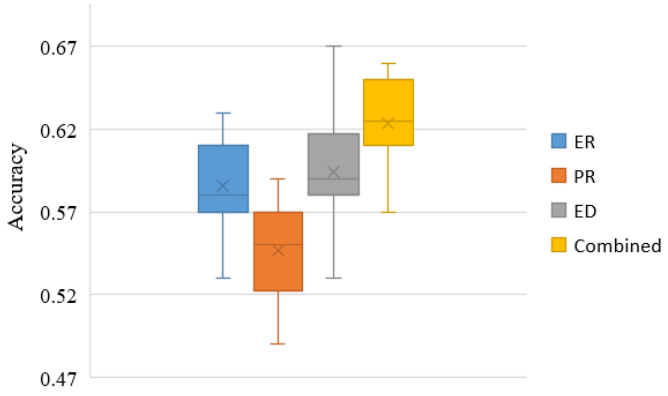

Fig. 3. Distribution of detection performances across all evaluated models for both individual and combined tasks. The considered features are fused at the feature level.

sense. In addition, the impact of segment lengths on the detection performance is minimal for the combined case.

C. Evaluation on the test set

The number of allowed submissions for test set predictions to the challenge is restricted to three. Table V shows the detection performances for all three attempts as well as for two created baselines by the challenge organizers. The 1^{st} attempt corresponds to the best-performing detector in Table III. The 2^{nd} and 3^{rd} attempts correspond to the best-performing detector developed on MFCC+AVEC-16 and eGeMAPS+MFCC+AVEC-16 feature sets, respectively, in Table IV. From Table V, it is evident that all three of our submissions surpassed the baseline established using eGeMAPS+SVM in terms of the accuracy metric.

TABLE V
SUICIDAL RISK DETECTION PERFORMANCES ON THE TEST SET.

Attempt	Accuracy	Precision	Recall	F1-score
Baseline*	0.51	0.51	0.58	0.54
Baseline†	0.61	0.61	0.60	0.61
1^{st}	0.54	0.53	0.66	0.59
2^{nd}	0.57	0.56	0.62	0.59
3^{rd}	0.55	0.54	0.68	0.60

$^{*}eGeMAPS + SVM$; $^{\dagger}W2V2 + BERT$

However, they did not outperform the baseline created with W2V2+BERT. Additionally, it is noteworthy that all three of our submissions achieved higher performance than both baselines in terms of the recall metric. According to the literature [22], the recall metric is preferred over the accuracy metric for classifying suicidal risk. This is because, in high-risk scenarios like suicidal risk assessment, failing to identify a suicidal case (false negative) can have serious consequences. Recall helps minimize such risks by ensuring that the majority of at-risk individuals are correctly identified. Thus, our proposed method is better than the provided baselines in the overall sense.

VI. CONCLUSIONS

This study presents a light-weight suicidal risk detection system. The database utilized in this study is provided in the 1st SpeechWellness Challenge organized in 2025. It comprises participants' spoken responses to three distinct tasks. The created baselines by the challenge organizers involve conventional machine learning and deep learning based binary classifiers. These classifiers are constructed either using data from a single task or by aggregating the predictions of three separate classifiers – each trained on data from one of the three tasks – through the majority voting (MV) approach. The results show that upon combining the three classifiers, the detection performance degrades in most cases compared to the detection performance of individual classifiers. We hypothesized that this degradation is due to the significant difference in the duration of spoken responses across three tasks. To tackle this problem, we proposed an alternative method in which the MV approach is applied twice instead of once. The MV approach is applied at the decision level of each task as well as on the accumulated decisions of three tasks. In addition, we explore three distinct hand-crafted feature sets and their fusion at the decision and feature levels for detecting suicidal risk. The efficacy of five different input segment lengths varying from 2 to 10 seconds in steps of 2 seconds is also analyzed. The results show that our proposed method outperformed both baselines on the development set in terms of accuracy. The best increment in the detection performance is noted to be 8%. On the test set, our submissions outperformed the baseline created using conventional machine learning only in terms of the accuracy metric. However, they did outperform both baselines in terms of recall metric, which is considered the most critical metric for high-risk scenarios like suicidal risk

assessment [22]. Thus, our proposed method is better than the provided baselines in the overall sense.

REFERENCES

[1] J. J. Gross, "The emerging field of emotion regulation: An integrative review," *Review of General Psychology*, vol. 2, no. 3, pp. 271–299, 1998.

[2] E. Rogante, M. Cifrodelli, S. Sarubbi, A. Costanza, D. Erbuto, I. Berardelli, and M. Pompili, "The role of emotion dysregulation in understanding suicide risk: A systematic review of the literature," *Healthcare*, vol. 12, no. 2, p. 169, 2024.

[3] W. Wu, Z. Cui, C. Lei, Y. Duan, D. Qu, J. Wu, B. Zhou, R. Chen, and C. Zhang, "The 1st speechwellness challenge: Detecting suicide risk among adolescents," *arXiv preprint arXiv:2501.06474*, 2025.

[4] S. Scherer, J. Pestian, and L.-P. Morency, "Investigating the speech characteristics of suicidal adolescents," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 709–713.

[5] V. Venek, S. Scherer, L.-P. Morency, A. Rizzo, and J. Pestian, "Adolescent suicidal risk assessment in clinician-patient interaction: A study of verbal and acoustic behaviors," in *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 277–282.

[6] J. Gideon, H. T. Schatten, M. G. McInnis, and E. M. Provost, "Emotion recognition from natural phone conversations in individuals with and without recent suicidal ideation," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 3282–3286.

[7] S. N. Chakravarthula, M. Nasir, S.-Y. Tseng, H. Li, T. J. Park, B. Baucom, C. J. Bryan, S. Narayanan, and P. Georgiou, "Automatic prediction of suicidal risk in military couples using multimodal interaction cues from couples conversations," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 6539–6543.

[8] D. J. France, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 47, pp. 829–837, July 2000.

[9] A. Ozdas, R. Shiavi, S. Silverman, M. Silverman, and D. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 9, pp. 1530–1540, 2004.

[10] S. Scherer, J. Pestian, and L.-P. Morency, "Investigating the speech characteristics of suicidal adolescents," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 709–713.

[11] H. K. Keskinpala, T. Yingthawornsuk, D. M. Wilkes, R. G. Shiavi, and R. M. Salomon, "Screening for high risk suicidal states using mel-cepstral coefficients and energy in frequency bands," in *2007 15th European Signal Processing Conference*, 2007, pp. 2229–2233.

[12] A. Belouali, S. Gupta, V. Sourirajan, J. Pestian, and R. Norel, "Acoustic and language analysis of speech for suicidal ideation among us veterans," *BioData Mining*, vol. 14, no. 11, 2021.

[13] V. Venek, S. Scherer, L.-P. Morency, A. izzo, and J. Pestian, "Adolescent suicidal risk assessment in clinician-patient interaction," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 204–215, 2017.

[14] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[15] Z. Cui, C. Lei, W. Wu, Y. Duan, D. Qu, J. Wu, R. Chen, and C. Zhang, "Spontaneous speech-based suicide risk detection using whisper and large language models," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2024, pp. 2915–2919.

[16] M. Gerczuk, S. Amiriparian, J. Lutz, W. Strube, I. Papazova, A. Hasan, and B. W. Schuller, "Exploring gender-specific speech patterns in automatic suicide risk assessment," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2024, pp. 1095–1099.

[17] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[18] Z. Cui, C. Lei, W. Wu, Y. Duan, D. Qu, J. Wu, R. Chen, and C. Zhang, "Spontaneous speech-based suicide risk detection using whisper and large language models," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2024, pp. 2915–2919.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[20] D. V. Sheehan, Y. Lecrubier, K. H. Sheehan, P. Amorim, J. Janavs, E. Weiller, T. Hergueta, R. Baker, G. C. Dunbar *et al.*, "The mini-international neuropsychiatric interview (mini): the development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10," *Journal of Clinical Psychiatry*, vol. 59, no. 20, pp. 22–33, 1998.

[21] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 1483–1484.

[22] C. G. Walsh, J. D. Ribeiro, and J. C. Franklin, "Predicting risk of suicide attempts over time through machine learning," *Clinical Psychological Science*, vol. 5, no. 3, pp. 457–469, 2017.