Psychological Crisis Detection in Crisis Hotline Chats Based on Weighted-Bidirectional Long Short-Term Neural Network*

Zhong Ding[1,2,3], Yang Zhou[2,4], Baoliang Zhong[2,4], Chenling Liu[1,2], and Zhentao Liu[2,3**]

[1] Institute of Education, China University of Geosciences, Wuhan 430074, China
[2] Psychological Science and Health Research Center,China University of Geosciences, Wuhan 430074, China
[3] School of Automation, China University of Geosciences, Wuhan 430074, China
[4] Wuhan Mental Health Center, Huazhong University of Science and Technology, Wuhan 430022,China

Abstract. Identifying crisis callers on psychological assistance hotlines is crucial for providing timely and effective support. To achieve this, the Crisis Psychological Hotline Corpus (CPHC) dataset was developed for suicidal crisis recognition. Statistical analysis reveals that non-crisis callers have significantly longer speech durations compared to operators, while there is no significant difference in speech duration between mild and severe crisis situations. The results show that the model achieves high accuracy and consistency in recognizing callers at different crisis levels, with an F1 score and accuracy of about 0.92 on the test set. Future research could expand the dataset, optimize the automated screening and identification system, and explore the detection of other mental health conditions to enhance the model's effectiveness in psychological assistance and suicide prevention.

Keywords: Psychological crisis detection· Suicide crisis· Affective computing· Weighted-BiLSTM.

1 Introduction

1.1 Background

Suicide is the most serious consequence of complex psychological struggles and has become a significant public health problem worldwide [1]. According to the World Health Organization, it remains one of the leading causes of death globally,

* Supported by the National Natural Science Foundation of China under Grant 61976197, and Grant 61403422;in part by the 111 Project, China, under Grant B17040; and in part by the College Students' Launching Project of Independent Innovation Funding Progam, China University of Geosciences (Wuhan) under Grant CUGDCJJ202246.

** Corresponding author: liuzhentao@cug.edu.cn

with over 700,000 people dying by suicide each year and many others attempting it [2]. Suicide not only causes irreversible damage to individuals but also has a lasting negative impact on families, friends, communities, and entire nations [3].

Suicide is not a sudden event but a gradual and cumulative result of worsening psychological states. Research has shown that psychological assistance hotlines are widely used worldwide as a crucial tool for suicide prevention and intervention [4]. These hotlines have proven valuable in managing psychological stress and providing crisis intervention for high-risk callers due to their timeliness, anonymity, user autonomy, affordability, and convenience [5].

However, as the workload of psychological assistance hotlines continues to increase, the shortage of professional operators and the prevalence of burnout have become more pronounced. This situation leads to an inability to provide optimal responses to crisis callers, thereby compromising the effectiveness of professional crisis intervention [6]. Data indicates that psychiatric and psychological issues account for approximately 30% of calls to psychological assistance hotlines [7]. This disparity between the demand for hotline services and the available operator capacity underscores the urgent need for new technological tools to support operators in managing the growing volume of calls.

In this context, automated crisis recognition systems are particularly important. By analyzing callers' voice characteristics, these systems can provide early warnings of crisis situations, enhancing hotline service effectiveness. This helps operators support high-risk callers more efficiently and offers reliable assistance in case of operator burnout.

1.2 Related work

With the development of affective computing and physiological-psychological computing, an increasing number of studies are exploring the detection of emotional and psychological problems through speech signals. Significant progress has been made in this area. Terra et al. suggest that detecting emotions, depression, and other psychological issues through speech has shown considerable potential [8].

Deep learning methods have been widely used in the field of emotion recognition. Meena et al. utilized deep learning techniques to recognize and classify emotions in speech, achieving an accuracy of 79% to 95% by classifying seven categories of emotions on two publicly available datasets [9]. Additionally, Wu et al. conducted a study on depression detection through audio information, achieving significant experimental results with a high performance score of 0.98 on independent test data from a public dataset [10].

For suicide risk detection, Belouali et al. characterized the speech of U.S. veterans and used the random forest algorithm for suicide detection, achieving 86% sensitivity and 70% specificity [11]. Speech research in crisis psychological hotlines is still in the exploratory stage, with particular attention paid to crisis detection. Grimland et al. noted that some risk factors in psychological crisis hotlines, such as displays of despair and crying, were highly correlated with suicide risk [12]. A study on the psychological hotline of the U.S. Department of

Veterans Affairs showed that machine learning-based speech distress detection is promising, and the proposed method achieved an AUC result of 0.86 in the experiment [13].

However, recognizing suicidal crisis and risk of psychological hotline callers in practical applications still requires a large amount of data and more accurate models. To address this need, this paper builds a hotline speech dataset specifically for suicidal crisis recognition and proposes a deep learning-based model to enhance the accuracy of suicidal crisis detection.

2 Crisis Psychological Hotline Corpus(CPHC)

Training and evaluating our proposed crisis recognition network requires the creation of a dataset. The Crisis Psychological Hotline Corpus, a crisis hotline speech dataset, was created by organizing and analyzing the speech source data. The dataset will not be made publicly available due to participant privacy protection. However, we are willing to provide a license with additional confidentiality conditions to interested researchers.

2.1 Crisis Psychological Hotline Source Data

This dataset was built from real speech dialogue sources. After obtaining permission from the Ethics Committee of the Wuhan Mental Health Center, affiliated with Huazhong University of Science and Technology, we collected 81 authentic voice recordings from the Mindfulness Crisis Intervention Hotline at the Wuhan Mental Health Center. These recordings, ranging from 5 to 65 minutes, captured the dialogues between the callers and the operators. To protect the privacy of both callers and operators, all personal information and identifiable features were removed from the audio.

The data were labeled using the Modified Suicide Risk Scale (SRS), an observational checklist for researching recent changes in suicidal ideation during telephone dialogues [4]. Gould et al. developed three subscales based on a literature review and input from telephone crisis workers: willingness to die, despair, and psychological distress, which serve as primary outcomes. Each subscale was assessed by two inventory items, with each item scored on a five-point scale: not at all, a little, average, quite a bit, and very much. Higher scores indicate a greater urgency to commit suicide. Shaw et al. demonstrated the reliability of the Modified Suicide Risk Scale as a standard for assessing caller suicide risk in an Asian population [14].

Specific labeling was conducted by five professionals with backgrounds in psychiatry and psychology. First, two graduate psychology students were trained in the use of the SRS and then practiced it. The agreement statistic (Cohen's kappa) between the two evaluators was higher than 90% [15] in order to start independent annotation of the source data audio based on the SRS. Sections where disagreement existed were secondarily evaluated by psychology professors and psychiatrists. The annotations with no disagreement were adopted, and those

Table 1. Crisis Psychological Hotline Source Data

Category	Number of audio samples			Audio duration (min)			
	Male calls	Female calls	Total	Male calls	Femal calls	Total duration	Mean duration
Acute Suicide Calls	4	5	9	184	160	345	38.3
Suicide calls	6	5	11	172	119	291	26.5
Crisis calls	10	10	20	356	279	636	31.8
Non-suicide calls	30	31	61	742	812	1555	25.5
Total	40	41	81	1098	937	2191	27.1

that remained in disagreement were independently assessed again by a third expert to ensure the quality of the assessment. After a careful labeling process, 9 call recordings were labeled as acute suicidal calls and 11 call recordings were labeled as suicidal calls. Both acute suicidal calls and suicidal calls were counted as crisis calls, and the remaining 61 calls were labeled as non-suicidal calls. The status of the collated counseling hotline source data is shown in Table 1.

2.2 Creation of the dataset

On the basis of the source data, the speaker sources of the source data were segmented.

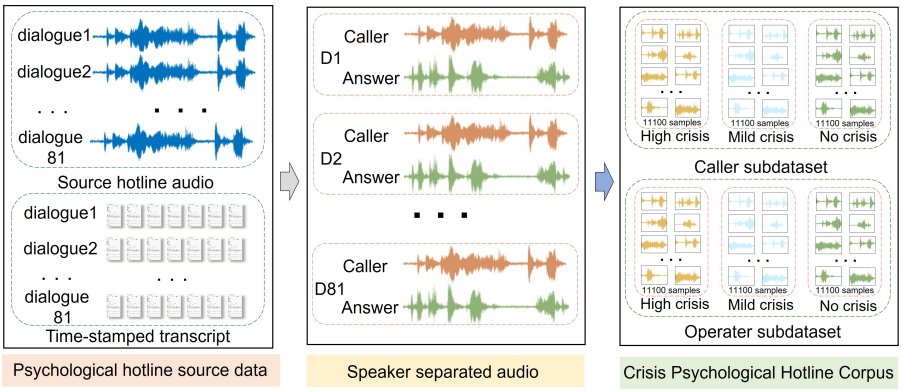


Fig. 1. Source data processing and dataset creation.

Each phone call recording was segmented into caller speech and operator speech according to the timestamp of the manually transcribed text. After removing silent segments and poor-quality audio segments, the remaining speech was segmented into speech segments of 5 seconds in length with a sampling rate of 16 kHz as a way to build a specialized speech dataset[16], and the specific steps of implementation and operation are shown in Figure 1.

Table 2. Distribution of self-constructed CPHC

Category	Speaker classification duration (mins)			Number of speaker classification segments		
	Caller	Operater	Total	Caller	Operater	Total
Suicide Calls	2	361	636	2749	4184	6943
Acute Suicide calls	1	193	345	1495	2260	3755
Crisis calls	1	168	291	1254	1924	3178
Non-suicide calls	8	657	1555	10444	7225	17669
Total	12	1379	2827	15942	15593	31535

2.3 Statistical analysis of the dataset

The dataset and source data were analyzed with descriptive statistics[17], and t-tests[18] were performed for differences in voice duration across in different crisis situations.

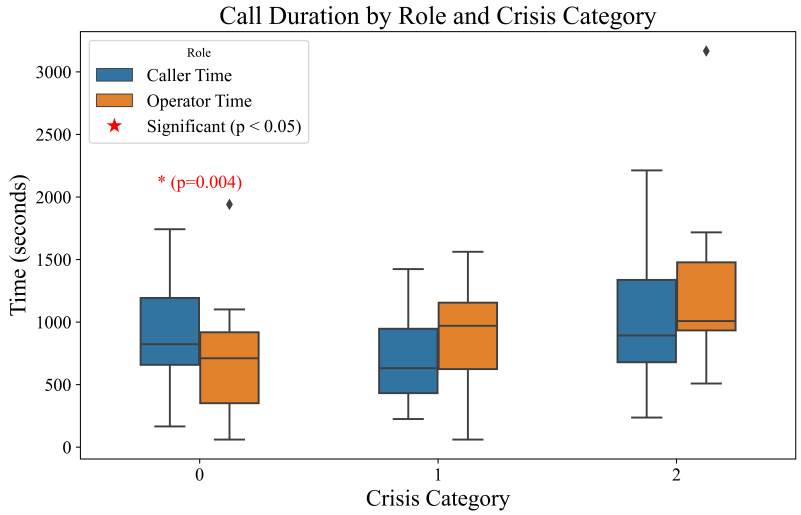


Fig. 2. Significance test for caller and operator hours in different categories of incoming calls.

In non-crisis situations (category 0), the caller's voice duration was significantly longer than the operator's voice duration, as shown in Figure 2. This indicates that in the absence of an apparent crisis, callers tend to spend more time expressing their situation and needs, while operators listen and record more. However, in mild and severe crisis situations (categories 1 and 2), there is no significant difference in voice duration between operators and callers. This is likely

because both parties need more interaction and communication when facing an actual crisis.

These findings provide a theoretical basis for automated screening and crisis identification that can take gender differences into account when designing psychological assistance and crisis intervention systems to provide more accurate and individualized support. Future research could further optimize the automated screening and identification system, delve into the psychological and social factors behind gender differences, and expand the size of the dataset to validate the generalizability and accuracy of the model.

3 Psychological Crisis Detection Modeling for CPHC

The method proposed in this paper is based on the bidirectional long and short-term memory network (BiLSTM)[19] commonly used in speech emotion recognition (SER)[20], which is capable of capturing both contextual and time-series signals of speech. Considering the specificity of the task of recognizing speech in psychological crisis, a weighing module is added to the original BiLSTM in order to improve the model's accuracy in recognizing a few classes (crisis speech), with the composite multiple speech features, such as MFCC and Mel cepstrum features, are used to maximize the retention of potentially effective features in speech. We believe that the proposed method can provide adequate performance performance.The proposed model is shown in Figure 3.

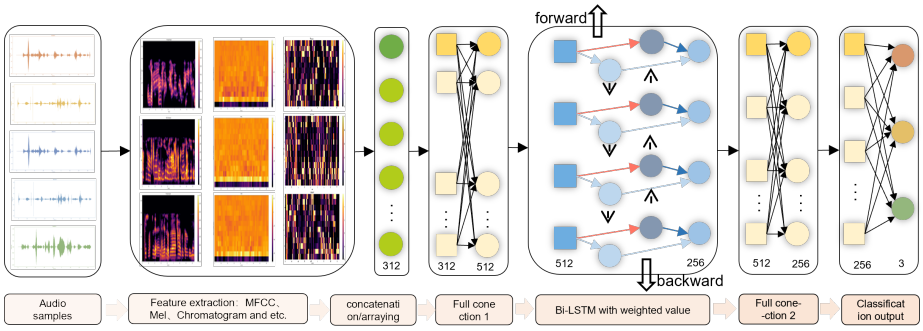


Fig. 3. Framework of psychological crisis detection.

3.1 Speech Feature Extraction

On top of the dataset, the audio was manipulated through preprocessing and the individual features extracted were concatenated, totaling 312 dimensional features as shown in Table 3.

Multiple sets of speech and acoustic functional features were extracted using Librosa[21]. First, we computed spectrograms using the Short-Time Fourier

Table 3. Extracting features and dimensions

Features	Dimensions
Spectral features (center of mass + plane)	3+1
MFCC features	50
Mel spectrogram	128
Spectral amplitude and rms energy	3+3
Chromatogram	12
Pitch	3
Spectral Contrast and Over Zero Rate	7+1
Total	312

Transform (STFT) and extracted pitch (pitch) and corresponding magnitude (magnitude) from them. Then, we extracted several statistical features, including spectral center of mass, spectral plane, MFCC (50 dimensions), chromatogram, Mayer spectrogram, spectral contrast, and over-zero rate. These features cover a wide range of attributes in both the time and frequency domains, which together form a comprehensive set of acoustic features. In addition the amplitude characteristics of the audio signal were obtained by calculating the spectral amplitude and the root mean square energy (RMS).

3.2 Weighted-BiLSTM Network

This study proposed a deep learning model specialized for speech classification, which is designed to cope with the psychological crisis speech recognition task. The model first extracts a variety of audio features, including Mel Frequency Cepstrum Coefficients (MFCC), Mel Spectrogram and Chromaticity Map. These features are merged and fed into a primary fully connected layer containing 512 neurons for initial feature analysis.

To effectively capture the temporal dynamics of audio data and to address the limitations of traditional models in dealing with category-imbalanced data, the model employs a bi-directional long- and short-term memory network with weights (Weighted-BiLSTM). This high-level component is crucial not only for managing the time series of audio signals, but also specifically enhances the model's ability to learn key categories, such as crisis speech, through a weighting mechanism that ensures effective recognition of a small number of categories. These features are then further processed through a second fully connected layer containing 256 neurons to optimize the depth of feature interpretation.

The ultimate phase of classification is performed through a layer containing three neurons, each representing a unique audio category: non-crisis calls, crisis calls, and urgent crisis calls. To address the category imbalance present in the dataset - an initial ratio of approximately 8:1:1, the model employs an integrated approach to data balancing. This included undersampling the major categories, oversampling the minority categories, and implementing a synthetic minority oversampling technique (SMOTE) to adjust the training sample distribution to a more equitable 2:1:1.

It is worth noting that these balancing techniques are only applied in the training phase to prevent distorting the natural data distribution in the validation phase. In addition, to meet the challenges of such computationally intensive tasks, CUDA techniques are employed to significantly speed up the processing time and make model training more efficient.

3.3 Experimental setup and performance metrics

A computer with a 12-core Intel Xeon Gold CPU (2.60 GHz) and an NVIDIA A6000 GPU with 48GB of memory was used to train the model. Pytorch 2.3.1 was chosen for the deep learning framework and experiments were conducted in a Python 3.8.2 environment.

To ensure the validity and reliability of the training, the experiments were conducted using the StratifiedKFold module for 5-fold cross-validation, with 80% as the training set and 20% as the validation set. StratifiedKFold is a stratified random sampling technique that preserves the proportion of samples from each category in the new training and validation sets, thus avoiding selection bias [22]. The experiments set up an early stopping mechanism, which stops training when the validation set performance metrics do not improve for 10 epochs in a row.

In addition, the weight values of different categories of samples are taken into account in the calculation of the loss function. The crisis recognition model is established mainly to accurately recognize crisis callers, so the weight values of different category classifications are set in the loss function rezaei2020addressing, in which the weights of general problem callers, suicide callers and acute suicide callers are 1.0, 2.0 and 2.0, respectively. The loss calculation formula is

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{2} w_c \cdot y_{i,c} \cdot \log(p_{i,c}) \tag{1}$$

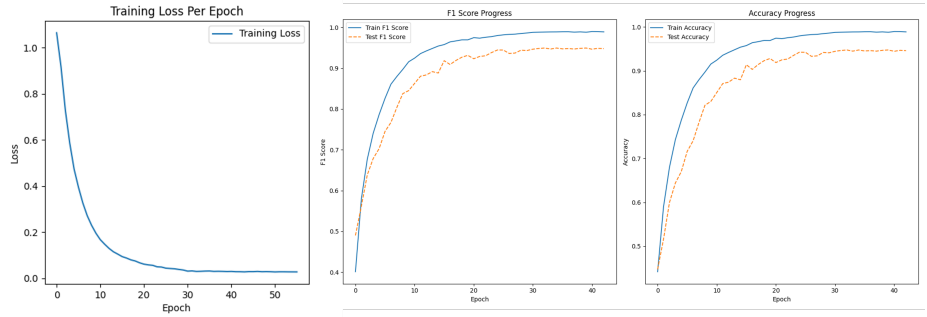
In the process of crisis speech recognition, it is very important to recognize the crisis caller. Therefore, both false-positive (false alarms) and false-negative (missed alarms) error cases need to be fully considered. False positives refer to the absence of a suicidal crisis being recognized as the presence of a suicidal crisis, and false negatives refer to the presence of a suicidal crisis itself being recognized as a non-crisis caller. Underreporting often has serious consequences and may result in missed opportunities to save lives because of failure to recognize a crisis in time[23]. At the same time, high false positive results can be financially and psychologically burdensome[24]. Considering these two factors, this study sets the F1 score as the main performance indicator to evaluate both its Accuracy and ROC curve results. The calculation formula of F1 score in this study is

$$\text{F1}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{2}$$

$$\text{F1}_{\text{weighted}} = \sum_{i=1}^{N} \frac{n_i}{n} \cdot \text{F1}_i \tag{3}$$

4 Results

In this study, we have achieved good results by proposing a deep learning model incorporating BiLSTM networks for automated recognition of suicide crisis callers using the self-constructed CPHC dataset.

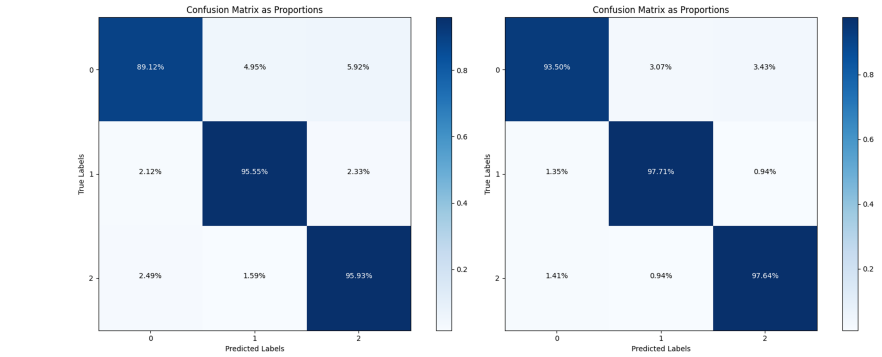


(a) training loss (b) performance progress

Fig. 4. Decline in training loss.

Figure 4-(a) illustrates the trend of the training loss of the model over epochs. It can be observed that the loss value gradually decreases and stabilizes as training proceeds, indicating that the model gradually converges without overfitting. The steady decrease of the training loss indicates that the model is constantly learning and optimizing, can adapt to the training data and has a strong generalization ability. Figure 4-(b) shows the progress of the model's F1 score and accuracy on the training and validation sets, and both the F1 score and accuracy gradually increase and eventually stabilize, and the F1 score and accuracy on the validation set reach about 0.97. This shows that the model not only can effectively identify the suicidal crisis callers, but also exhibits good stability and reliability during the training process.

From the confusion matrices presented in Figures 5, it can be seen that the model performs more consistently across folds, with the accuracy of non-crisis callers (category 0) ranging from 89.12% to 93.50%, the accuracy of mild crisis callers (category 1) ranging from 95.55% to 97.71%, and the accuracy of severe crisis callers (category 2) ranging from 95.93% to 97.64%. These results indicate that the model has high accuracy and consistency in recognizing mild and severe crisis callers, and is able to effectively differentiate between different levels of crisis callers. The high accuracy and consistency of the model gives it significant potential for practical application to assist operators in handling crisis callers more effectively. These results demonstrate that the proposed method has good performance in handling speech recognition tasks and is able to maintain consistency during training and validation.



(a) 5-fold worst performance (b) 5-fold best performance

Fig. 5. Best and worst performance in k-fold cross-validation.

Table 4. Compared methods

Methods	Test Accuracy(%)	Test F1-score(%)
SVM	79.4	81.3
XG-Boost	83.0	83.6
Bi-LSTM	87.7	88.6
Weighted-Bi-LSTM(our)	**92.1**	**92.5**

Table 4 presents a comparison of methods focusing on test accuracy and F1-score.Our Weighted-Bi-LSTM method surpassed all with 94% accuracy and 96% F1-score. These results highlight the high accuracy and F1 scores of our model, especially in recognizing mild and severe crisis callers. The success of our model is due to the Weighted-Bi-LSTM's ability to capture and prioritize important features, effectively leveraging temporal dependencies and relevant features, leading to superior performance.

5 Conclusion

This study has three main contributions, firstly we proposed a crisis identification dataset for psychological crisis assistance hotlines in the Chinese context, secondly we statistically analyzed it to get one important finding, and based on this we proposed a weighted-BiLSTM network for identifying crisis callers.

Statistical analysis of the dataset shows that in non-crisis situations, callers spoke longer than operators, but this difference disappeared in crisis situations. To the best of our knowledge, this is the first time that differences in psychological crisis help lines have been analyzed at the level of voice calls. These findings suggest new approaches for automated crisis recognition and highlight the value of suicide crisis identification for psychological hotlines, helping operators more efficiently identify high-risk callers and improve intervention effectiveness. The weighted-BiLSTM model proposed in this study, incorporating features like

MFCC and Mel cepstrum coefficients, effectively captures contextual and temporal speech signals. It demonstrated high accuracy in recognizing crisis callers using 5-fold cross-validation. To address data imbalance, we used undersampling, oversampling, and SMOTE techniques to ensure a 2:1:1 training set ratio. Adjusting class weights in the loss function further enhanced recognition performance.

Despite the model's strong performance, there are some limitations. The dataset size needs further expansion to enhance the model's generalizability and robustness. While data balancing techniques were employed, real-world applications may still face data imbalance issues, requiring further optimization. Future research should focus on optimizing the model and expanding the dataset to improve its generalizability and robustness. Additionally, future studies could explore the detection of other mental health issues related to changes in speech patterns, such as anxiety and PTSD. Integrating psychological knowledge with AI algorithms can further enhance the model's reliability. Collecting more diverse data will improve the model's adaptability and accuracy in various contexts.

These improvements will provide stronger technical support for crisis interventions in psychological assistance hotlines, ultimately aiding in suicide prevention. This will help mitigate the negative impact of suicide on individuals, families, and society, and improve public health. Through continuous optimization and expansion, we believe the model will have a significant impact in the field of mental health, providing robust support for crisis intervention.

References

1. Turecki, G., Brent, D.A., Gunnell, D., O'Connor, R.C., Oquendo, M.A., Pirkis, J., Stanley, B.H.: Suicide and suicide risk. Nature reviews Disease primers **5**(1), 74 (2019)
2. Organization, W.H., et al.: Suicide worldwide in 2019: global health estimates (2021)
3. Organization, W.H., et al.: Live life: an implementation guide for suicide prevention in countries (2021)
4. Gould, M.S., Kalafat, J., HarrisMunfakh, J.L., Kleinman, M.: An evaluation of crisis hotline outcomes. part 2: Suicidal callers. Suicide and Life-Threatening Behavior **37**(3), 338–352 (2007)
5. Ingram, S., Ringle, J.L., Hallstrom, K., Schill, D.E., Gohr, V.M., Thompson, R.W.: Coping with crisis across the lifespan: The role of a telephone hotline. Journal of Child and Family Studies **17**, 663–674 (2008)
6. Stamm, B.H.: Helping the helpers: Compassion satisfaction and compassion fatigue in self-care, management, and policy of suicide prevention hotlines. Resources for community suicide prevention pp. 1–4 (2012)
7. Jia, Q., Wu, Z., Li, Z., Zhang, X., Luo, L., et al.: Basic analysis of calls from suzhou psychological aid hotline from 2010 to 2020. Advances in Educational Technology and Psychology **5**(4), 73–80 (2021)
8. Terra, M., Baklola, M., Ali, S., El-Bastawisy, K.: Opportunities, applications, challenges and ethical implications of artificial intelligence in psychiatry: a narrative review. The Egyptian Journal of Neurology, Psychiatry and Neurosurgery **59**(1), 80 (2023)

9. Meena, G., Mohbey, K.K., Indian, A., Khan, M.Z., Kumar, S.: Identifying emotions from facial expressions using a deep convolutional neural network-based approach. Multimedia Tools and Applications **83**(6), 15711–15732 (2024)
10. Wu, P., Wang, R., Lin, H., Zhang, F., Tu, J., Sun, M.: Automatic depression recognition by intelligent speech signal processing: A systematic survey. CAAI Transactions on Intelligence Technology **8**(3), 701–711 (2023)
11. Belouali, A., Gupta, S., Sourirajan, V., Yu, J., Allen, N., Alaoui, A., Dutton, M.A., Reinhard, M.J.: Acoustic and language analysis of speech for suicidal ideation among us veterans. BioData mining **14**, 1–17 (2021)
12. Grimland, M., Benatov, J., Yeshayahu, H., Izmaylov, D., Segal, A., Gal, K., Levi-Belz, Y.: Predicting suicide risk in real-time crisis hotline chats integrating machine learning with psychological factors: Exploring the black box. Suicide and Life-Threatening Behavior (2024)
13. Pacula, M., Meltzer, T., Crystal, M., Srivastava, A., Marx, B.: Automatic detection of psychological distress indicators and severity assessment in crisis hotline conversations. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4863–4867. IEEE (2014)
14. Shaw, F.F.T., Chiang, W.H.: An evaluation of suicide prevention hotline results in taiwan: caller profiles and the effect on emotional distress and suicide risk. Journal of affective disorders **244**, 16–20 (2019)
15. Li, M., Gao, Q., Yu, T.: Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. BMC cancer **23**(1), 799 (2023)
16. Gaol, Y., Fernandez-Marques, J., Parcollet, T., de Gusmao, P.P., Lane, N.D.: Match to win: Analysing sequences lengths for efficient self-supervised learning in speech and audio. In: 2022 IEEE Spoken Language Technology Workshop (SLT). pp. 115–122. IEEE (2023)
17. Oh, D.M., Pyrczak, F.: Making sense of statistics: A conceptual overview. Routledge (2023)
18. Ross, A., Willson, V.L., Ross, A., Willson, V.L.: Paired samples t-test. Basic and Advanced Statistical Tests: Writing Results Sections and Creating Tables and Figures pp. 17–19 (2017)
19. Siami-Namini, S., Tavakoli, N., Namin, A.S.: The performance of lstm and bilstm in forecasting time series. In: 2019 IEEE International conference on big data (Big Data). pp. 3285–3292. IEEE (2019)
20. Madanian, S., Chen, T., Adeleye, O., Templeton, J.M., Poellabauer, C., Parry, D., Schneider, S.L.: Speech emotion recognition using machine learning—a systematic review. Intelligent systems with applications p. 200266 (2023)
21. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. In: SciPy. pp. 18–24 (2015)
22. Aoyama, H.: A study of stratified random sampling. Ann. Inst. Stat. Math **6**(1), 1–36 (1954)
23. Li, F., Yip, P.S.: How to make adjustments of underreporting of suicide by place, gender, and age in china? Social psychiatry and psychiatric epidemiology **55**, 1133–1143 (2020)
24. DeFrank, J.T., Barclay, C., Sheridan, S., Brewer, N.T., Gilliam, M., Moon, A.M., Rearick, W., Ziemer, C., Harris, R.: The psychological harms of screening: the evidence we have versus the evidence we need. Journal of general internal medicine **30**, 242–248 (2015)