2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus) | 978-1-6654-0476-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ElConRus51938.2021.9396540

The Detection of Depression Using Multimodal Models Based on Text and Voice Quality Features

Hanadi Solieman
Saint Petersburg Electrotechnical University "LETI"
St. Petersburg, Russia
khsoliman@stud.etu.ru

Evgenii A. Pustozerov
[1] Saint Petersburg Electrotechnical University "LETI"
[2] Almazov National Medical Research Centre
St. Petersburg, Russia
eapustozerov@etu.ru

***Abstract*—The article proves the concept that an automatic diagnosis of depression can be achieved using audio recordings of the individuals' voices. DAIC-WOZ database was used as a data source. Audio and textual data were preprocessed and converted to a set of optimized parameters for two models. Appropriate Deep Learning models to detect depression in the transcripts of the audio recordings and voice quality features, were utilized. We created a text analysis model on a word-level using Natural Language Processing (NLP) techniques, and a voice quality analysis model on tense to breathy dimension. The text analysis model made its best performance with an F1-score equal to 0.8 (0.42) for non-depressed (depressed) individuals, while the voice quality model scored 0.76 (0.38). As a result, we had two models that would be implemented in a system for the diagnosis of depression.**

Keywords*—*Depression; Deep Learning; text analysis; voice quality; semi-contextual; word-level; speaker-independent; DAIC-WOZ

I. Introduction

Depression is a common mental disorder. It is a leading cause of disability worldwide, a major contributor to the overall global burden of disease, and leads to suicide. More than 264 million people of all ages suffer from depression worldwide [1]. Unipolar depressive disorders are predicted to be the second leading cause of the burden of diseases by 2030 [2]. Stanislav Poltorak from the St. Petersburg V.M Bekhterev Psychoneurological Research Institute stated that 25% of the Russian population may be affected by depression [3]. According to the World Health Organization, almost 800 thousands people commit suicide every year, where it is considered to be the main cause of death for young people from 15 to 19 years old [4]. However, for each suicide, the cases of attempts are many more, WHO reported the number to be 20 attempts per suiside in 2014 [5]. In 2016, the suicide rate in the Russian Federation was 26,5 (per 100000 population) for both sexes but mostly men [6].

The subjectivity, inconsistency, and high-cost of the clinical diagnosis; make seeking help burdensome for undiagnosed depressed patients; especially in the early stages. Thus, the solution is to build an intelligent diagnosis system as referred to by telemedicine. Those systems are software and applications created from Machine Learning (ML) and Deep Learning models, which can initially analyze the patient evaluation to form a self-diagnosis [7], [8].

The aim of this work is the development of a Deep Learning model to detect depression without the need for a visit to the clinic.

II. Previous Work

A. The analysis of behavioral markers as indicators of depression

The behavioral markers of depression can be divided into verbal and non-verbal behaviors. The verbal behavior is connected with words and their meaning, while the non-verbal behavior varies between visual, acoustic, and anything without the use of language. Next, we will discuss these markers and the ability to use them by reviewing previous work and modern approaches related to the prediction of depression.

The study [9] indicated that patients with depression or anxiety behave differently from controls during conversations; they hold their heads even and tap their hands. However, motion tracking for those movements is highly complicated through activity units (AUs) alone. Besides, the signs of depression are better to be predicted from video-based features rather than frame-based ones [10], because the signs are temporal.

A lot of studies on depression implied that depressed people show less facial expressivity, such as the absence of a smile. Nevertheless, masked depression (smiling depression) is a type of depression with atypical syndromes, where individuals may seem perfectly happy and smiling [11], [12]. Thus, capturing the smile lines in facial features (2D or 3D) is not as effective as it seems for predicting depression against what is suggested in [13].

A comparison between different combinations of the three domains prosodics, glottal parameters, and the vocal tract, showed that a combination of prosodic and glottal feature had better results in indicating depression [14]. Moreover, the authors in [15] investigated both vocal jitter and glottal flow spectrum for indicating depression and near-term suicide. They suggested that analyzing the speech (audio recordings) of individuals by using their voice without content, can distinct between non-depressed and depressed patients. The study [16] aimed at analyzing the performance of using acoustic, glottal inverse filtered (GIF), and electroglottographic (EGG) signals to classify voice quality and the performance of different types of classifiers Support Vector Machines (SVM), Random Forest (RF), Deep Neural Network (DNN), and Gaussian Mixture

978-1-6654-0476-1/21/$31.00 ©2021 IEEE

Models (GMM). The 72 extracted features using the COVAREP toolbox were the conventional features.

The purpose of the study in [17] is to examine the ability to detect psychological distress (depression or post-traumatic stress disorder (PTSD)) in the early stages by the use of voice features to indicate tenseness in the participant's voice. Their analyzed dataset was DAIC-WOZ. Four features were used: the normalized amplitude quotient (NAQ), the quasi-open quotient (QOQ), the estimated open quotient based on the Mel frequency cepstral coefficients using Neural Network for the approximation (OQNN or ANN-OQ [18]), and the maxima of the slope of the regression line of the speech signal (peakSlope). SVMs were trained to classify the median of the features with the strategy of leave-one-out (leaving one sample out of the training process). The choice of the kernel type was the radial function, and the training process was repeated more than once. Performance of almost 75% was achieved, where the F1 score equals 0,769 for depressed and 0,727 for non-depressed. However, the cost of the estimation of OQNN in time and memory; which is done by training a Neural Network to approximate the values, make this feature overqualified to be used, and the issue of approximation is there. Further, by taking the median of the features to train the model, we will be losing most of the temporal information [19]. According to the study [20], the extracted slope coefficient from the proposed wavelet-based method; named "Peak Slope" was shown to be better in differentiating between breathy and tense voice quality and to be a robust parameter against high-level babble noise in the signal (SNR=10dB) with an accuracy of 75%. It is also mentioned that NAQ also showed robustness down to SNR=10 dB but with a lower accuracy (65%). While the Difference in the amplitude of the first two harmonics of the differentiated glottal source spectrum (H1-H2) parameter came third with accuracy near 63% and robustness to SNR=15dB. Furthermore, peakSlope is a standalone parameter, i.e. it does not need any other algorithms to be calculated. Moreover, it is obtained without assumptions.

Another study [21] on the use of the maxima dispersion quotient (MDQ) for the distinction between breathy and tense voices, revealed that MDQ improved the classification for voice quality. Also, it is shown that MDQ is robust to additive noise for SNR equal to 10 dB.

Based on a two-step feature selection method in [19], the gender feature and formant features were not selected as discriminative features, but COVAREP and AUs were most involved for 33 and 9 features respectively.

B. *Analyzing the model properties*

In this section, the modality and approach of the model are discussed. The modality of the model in our context is whether it processes one type of features; called unimodal, or process more than one type of features; multimodal. In the studies [16], [17], unimodal models were used to predict depression based on acoustic features. While in [19], [22], [23], both visual and acoustic features were involved in the training process. Moreover, a comparison between using a combination of textual and audio features or each type separately was made in [24]. The results indicated that using the combination of both; while applying data augmentation, had the best performance.

In the case of the approach, different studies are considered. In [19], the writers proposed an approach to detect depression in a voice recording based on an analysis of the subject's reaction to the context. They suggest that improvements can be made to the used method by mapping the subject's interview in the DAIC-WOZ database to a set (a vector) of features that maintain the context through the different sections of the interview. The first stage in this proposed method is to divide each interview into segments related to 83 topics, where only a few topics (i.e., 14 topics) can cover 80% of the interviews. A total of 46 features were used in a regressing Stochastic Gradient Descent (SGD) model with the best performance of root mean square error (RMSE) equals to 4,99 and F1-score equals to 0,6 on the test set. However, the topic segmentation is not automated which makes the approach unsuitable for generalization outside the scope of the PHQ interviews.

Despite that, the approach in [25] suggested that the detection of depression can be carried out full-automatically without the need for topic modeling, but instead by using sequential modeling. Learning from the sequence of questions and responses achieved as good as topic modeling using both audio and text features from DAIC-WOZ database. While the multi-modal sequence model of audio and text features performed the best. Audio and text features were processed separately, where bidirectional Long Short-Term Memory Neural Networks (LSTMs) were layered solely for each type of feature. A concatenation using feed-forward layers combined the results of both, to achieve an F1-score equal to 0,77.

In [26], the authors investigated the use of F2SVM in classifying depressed from non-depressed individuals based on voice quality features indicating the tense or breathy mode. The F2SVM outperformed the standard SVMs with an accuracy of 82% in frame-wise analysis, and 97% in sentence-wise. However, the input for this model is fuzzy labeled and the output is fuzzy too (muticlassified) not binary (0, 1), which would require conversion and arise some problems.

III. Database

A. *Description of the used database*

The Distress Analysis Interview Corpus (DAIC) is designed to support the diagnosis of psychological disorders, such as depression, anxiety, and PTSD [27]. This database is a set of conducted interviews with subjects who were identified as cases (diseased) and controls (healthy) regarding one or all of the previously mentioned disorders. The database includes recordings of multimodal dyadic interactions and information about the participants' condition based on the PHQ-8 questionnaire.

In the Wizard-of-Oz set (DAIC-WOZ), the interviewer is a computer agent, which conducts interviews to make the subject feels at ease to share information. This virtual interviewer tries to assess the indicators of distress disorders [28]. The animated virtual interviewer called Ellie is controlled by two human interviewers (wizards) in another room. The participant was alone in a room in front of a large computer screen. Ellie was designed to interact in favor of capturing verbal and non-verbal indicators (as shown in the video [29]) responsible for the assessment of distress disorders, such as depression.

B. Data-related problems

Due to the small sample size in datasets, it is highly required to analyze a small number of features to avoid exposing the model to overfitting and dimensionality problems. A possible way to avoid overfitting is by dropping samples during training [30]. This way is easily done by creating a layer called a Drop-out inside your Deep Learning model. Another way is early stopping, where we limit the number of trials (epochs) to the performance of the model at each trial, i.e. we stop training when we obtain satisfying results.

As for the dimensionality problem, it can be solved by using Transfer Learning. This technique is built to take the features learned for a problem and use them on another similar problem [31]. Thus, no need to train our model from scratch on the small number of samples that we have.

Data-imbalance (class-imbalance) is a common problem in most of the available datasets. In general, class-imbalance affect the performance of the model negatively by reducing the accuracy, i.e. increasing the error. There are ways to solve this problem on both the model-level and data-level. The model-level solutions include either changing the algorithm of the model or changing the performance metrics (error evaluation methods), where instead of using accuracy, we may use confusion matrix, precision, recall, or F1-score, this case is well known as the accuracy paradox [32]. In 2017, the ICCV award-winning paper [33] presented a reshaped cross-entropy loss function named Focal Loss, which decreases the weights for the samples in the majority class while focuses on the samples of the minority-class [34].

On the other hand, solutions on the data-level include resampling techniques and data augmentation. Despite that, a lot of studies chose to take a smaller number of samples, where both depressed and non-depressed samples are equal in number [35]–[37].

C. The used sets of data

Based on our analysis of previous studies, we decided to use the transcript of the interview to feed the text analysis model, while a set of audio features (NAQ, QOQ, H1-H2, MDQ, and peakSlope) from the COVAREP files to feed the voice quality model.

IV. Model Modality and Approach

A sequential multimodal model that can predict depression using two types of data was built. Regarding the use of context, a semi-contextual approach was used, where NLP techniques were exploited to perform contextual prediction on the word-level. We named it 'semi' because it doesn't perform on the level of the contextual topic. The two chosen data types are text and audio features. Each model has its strategy and can be considered to be a separate unimodal model. The first unimodal model is designed to accomplish the task of text analysis of the participants' interview transcripts. On the other hand, the second unimodal model is responsible for voice quality analysis of the extracted audio features from the participants' voices during the interviews.

A. Text analysis model

The text analysis model processes the transcripts of the participants' interviews with Ellie using NLP techniques. It deploys just one type of linguistic cue which is sentiment analysis while ignores the other two types represented by the names syntactic structure and lexical features analysis. The model processes the responses of the participants on the word-level, where it analyzes the semantic content of each word by creating vectors of related words using embeddings. The model is language-dependent, where the transcripts are in English; as well as the source of the transferred word-vector weights. The first stage to train the model is data loading. The second stage is the preprocessing of the loaded data. Then comes the construction of the model.

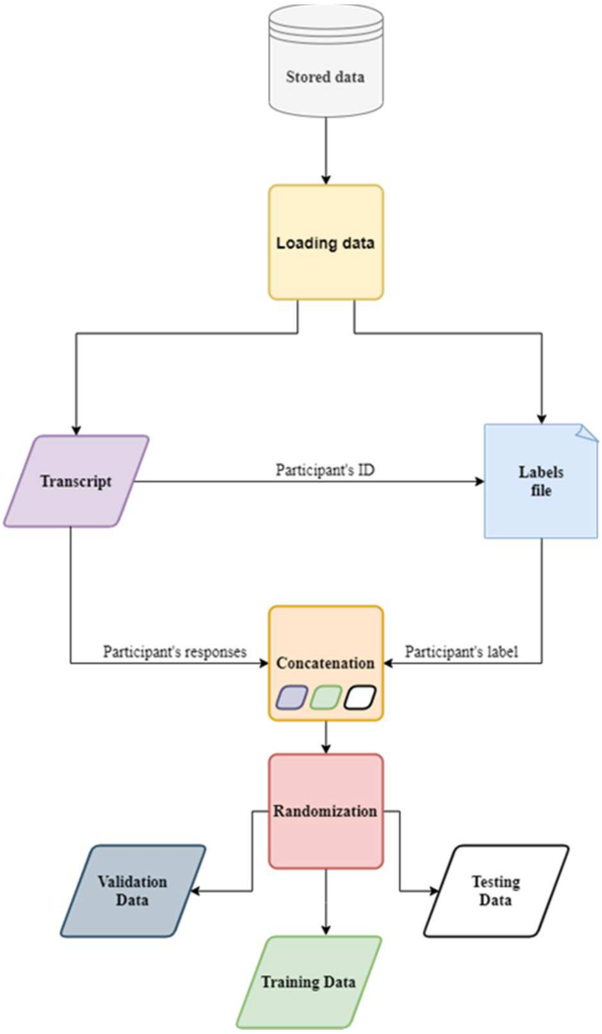

Fig. 1. The diagram of data loading for the text analysis model

In fig. 1, we illustrate the different steps of data loading for the text analysis. We start by loading the data from the storage on the cloud service. We extract the participant's ID from the transcript file name to match it with its label from the labels file. Then only the participant's responses were extracted through the whole interview and put in a list that starts with the label and belongs to one set of the following: training, validation, and testing according to the configuration of the database. The splitting is indicated in the diagram by placing small data boxes inside the box of the concatenation process. We did not draw them as independent entities because they are not the final results of the data loading stage. After putting all of the transcripts in lists with the labels, we apply randomization to avoid the possibility of the presence of patterns in the order of the data for each set. At the end of this stage, we get the final version of data before preprocessing, as three sets are named according to their purposes.

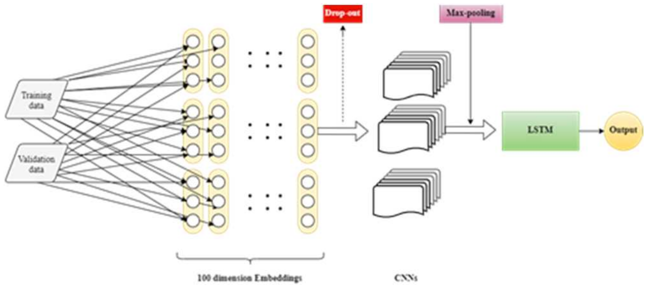

Fig. 2. The architecture of the text analysis model

The text analysis model is a sequential supervised learning model. Its input are sequences of values that will get processed by a set of sequential layers. This set of layers are built layer by layer without being linked to each other or shared. As illustrated in fig. 2, this unimodal model consists of one embedding layer with a dimension of 100, 200, or 300. We chose to draw the 100-dimensional embeddings layer in the figure below because it had the best results according to the related tests. The weights of the pre-trained word vectors are planted in this layer to increase the model capacity. Those weights are kept static (untrainable) so they will not be added to our parameters and violate our purpose on training. Then a drop-out layer is placed, to avoid the overfitting on the training data. The reminding parameters are fed to the CNN layer to extract more features out of them. The output of the CNN layer is summarized using a maximum pooling layer. Then it is fed to the LSTM neural network.. Finally, the output is formed in the dense layer to be a unique 1-dimensional decision on the participant's state as depressed or non-depressed.

B. *Voice quality model*

The voice quality analysis model processes the audio features of the participants' voices during the interviews with Ellie on a tense or breathy dimension. It deploys a set of audio features that are provided in the database using the COVAREP toolbox. The model is speaker-independent, where these features are not affected by the speaker's voice characteristics. The first stage to train this model is data loading. This stage differs from the data loading of the text analysis model by the addition of extracting the start and stop time of the participant's speech and ignoring the interviewer's and taking just one minute of speech for each participant. Then comes the construction of the model, where the voice quality analysis model is also a sequential supervised learning model. Its input is sequences of values that will get processed by a set of sequential layers. This set of layers are built layer by layer without being linked to each other or shared. The main difference between this model and the text analysis model is the absence of the embedding layer here, while the reminding layers are still the same. The input is fed to the CNN layer to extract features out of it. The output of the CNN layer is summarized using a maximum pooling layer. After that, it enters several LSTM layers, the number and the type of these layers are investigated. However, we drew it in the figure below as one LSTM layer, as it had the best performance. Finally, the output is formed in the dense layer with one node to be a unique 1-dimensional decision to determine the participant is depressed or not.

C. *Loss function and optimization*

The focal loss function was introduced for the classification tasks of imbalanced datasets, where it can concentrate on the hardly predicted class (minority class). It is a transformed form of the cross-entropy loss function. The scaled version of the focal loss function can be computed using 2.14:

$$FL(p_t) = -\alpha \times (1 - p_t)^{\gamma} \times \log{(p_t)} \qquad (1)$$

Where $FL(p_t)$ is the focal loss of the prediction p at the time t, α is the modulating (balancing) factor between the classes 0 (negative diagnosis), and 1 (positive diagnosis), and γ is the focusing factor which permits us to focus on the hard-to-predict class.

To ensure that our models are somewhere near their best performance, we apply optimizers on them while learning to get the best out of it. Based on previous work in the same field, we decided to investigate the use of two optimizers Adam and NAdam.

V. Code Configurations

In the implementation of this work, we used the latest stable releases of each one of the following: Python (3.7), open-source library TensorFlow (2.2.0), the Deep Learning Application Programming Interface (API) named Keras (2.3.0), and NumPy package. Also, Google Colab was used as a cloud service.

VI. Results

After conducting several experiments on the available options, the focal loss has better performance than binary cross-entropy regarding the minimization of the loss. Adam optimizer makes better results than NAdam for our data. Our results were using the test set, where we believe it's better to work on a set that the model did not see before. All our results scored better for the non-depressed class because it is the majority class. Remarkably, it performed better on the five glottal flow features (NAQ, QOQ, H1-H2, MDQ, and peakSlope) in comparison with the whole set of COVAREP. This contradicts the previous results of multiple studies. Still, our optimization was on just the 5 features, and that makes the results applicable just for it than any other set.

The best results of both models on the testing data are described in the table below.

Table. The Rsults on Test Data

Model	Focal Loss	Accuracy	Class	Precision	Recall	F1-score
Text analysis	0.35	0.70	Nondepressed	0.74	0.88	0.81

			Depressed	0.50	0.29	0.36
Voice quality	0.67	0.66	Nondepressed	0.77	0.73	0.75
			Depressed	0.44	0.50	0.47
Concatenation	0.07	0.68	Nondepressed	0.74	0.85	0.79
			Depressed	0.44	0.29	0.35

VII. Conclusion

The purpose of this study was achieved by developing data-driven models that can detect depression using audio recordings. In this research, all of the objectives were reached where two unimodal sequential models were developed and studied. We created a semi-contextual text analysis model using NLP techniques and deploying transfer learning. This model performs detection of depression on word-level using a transcript of the individual's interview with a virtual interviewer. Then we created a voice quality analysis model that uses five glottal flow voice features. It is a speaker-independent model that indicates depression by capturing the tenseness and breathiness in the individual's voice during the same interview. The used programming language to code both of those models is Python, implemented on Google Colab. Both models used focal loss function in a try to overcome the imbalance problem in the DAIC-WOZ database. The text analysis model detected depression with an F1-score equal to 0,8 on the non-depressed group of the test set. While the voice quality model made it with 0,75. They can be used after improvements as a self-diagnostic tool for depression without seeking clinical help. This study will treat the point of the automation of the diagnosis of depression, and pioneer for the automated detection of other mental illnesses.

References

[1] Depression. Available at: https://www.who.int/news-room/fact-sheets/detail/depression (Accessed 30 November 2020).

[2] Mathers C.D, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med., 2006, vol. 3, no. 11, p. e442 DOI: 10.1371/journal.PMED.0030442.

[3] Depression in Russia (2017). Available at: https://www.globallyminded.org/home/depression-in-russia/ (Accessed 30 November 2020).

[4] Suicide. Available at: https://www.who.int/news-room/fact-sheets/detail/suicide (Accessed 30 November 2020).

[5] Preventing suicide: A global imperative (2014). Available at: http://www.who.int/mental_health/suicide-prevention/world_report_2014/en/ (Accessed 30 November 2020).

[6] Suicide rate estimates, age-standardized - Estimates by country (2018). Available at: https://apps.who.int/gho/data/node.main.MHSUICIDEASDR?lang=en (Accessed 30 November 2020).

[7] Telemedicine: opportunities and developments in Member States: report on the second global survey on eHealth (2010). Available at: https://apps.who.int/iris/handle/10665/44497 (Accessed 30 November 2020).

[8] Pacis D. M. M., Subido E. D. C., Bugtai N. T. Trends in telemedicine utilizing artificial intelligence. AIP Conference Proceedings, 2018, vol. 1933, no. 1, p. 040009 DOI: 10.1063/1.5023979.

[9] Fairbanks L. A., McGuire M. T., Harris C. J. Nonverbal interaction of patients and therapists during psychiatric interviews. Journal of Abnormal Psychology, 1982, vol. 91, no. 2, pp. 109–119 DOI: 10.1037/0021-843X.91.2.109.

[10] Pampouchidou A., Simantiraki O., Fazlollahi A., Pediaditis M., Manousos D., Roniotis A., Giannakakis G., Meriaudeau F., Simos P., Marias K., Yang F., Tsiknakis M. Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text. Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC '16). Amsterdam, 2016, pp. 27–34 DOI: 10.1145/2988257.2988266.

[11] Bhattacharya S., Hoedebecke K., Sharma N., Gokdemir O., Singh A. "Smiling depression" (an emerging threat): Let's Talk. Indian Journal of Community Health, 2019, vol. 31, no. 04, p. 433–436. Available at: https://www.iapsmupuk.org/journal/index.php/IJCH/article/view/1255 (Accessed 30 November 2020).

[12] Lesse S. The masked depression syndrome: Results of a seventeen-year clinical study. American Journal of Psychotherapy, 1983, vol. 37, no. 4, pp. 456–475.

[13] Namboodiri S. P., Venkataraman D. A computer vision based image processing system for depression detection among students for counseling Indonesian Journal of Electrical Engineering and Computer Science, 2019, vol. 14, no. 1, pp. 503–512 DOI: 10.11591/ijeecs.v14.i1.pp503-512.

[14] Moore II E., Clements M. A., Peifer J. W., Weisser L. Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech. IEEE Transactions on Biomedical Engineering, 2008, vol. 55, no. 1, pp. 96–107 DOI: 10.1109/TBME.2007.900562.

[15] Ozdas A., Shiavi R. G., Silverman S. E., Silverman M. K., Wilkes D. M. Investigation of Vocal Jitter and Glottal Flow Spectrum as Possible Cues for Depression and Near-Term Suicidal Risk. IEEE Trans. Biomed. Eng., 2004, vol. 51, no. 9, pp. 1530–1540 DOI: 10.1109/TBME.2004.827544.

[16] Borsky M., Mehta D. D., Van Stan J. H., Gudnason J. Modal and Nonmodal Voice Quality Classification Using Acoustic and Electroglottographic Features IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, vol. 25, no. 12, pp. 2281–2291 DOI: 10.1109/TASLP.2017.2759002.

[17] Scherer S., Stratou G., Gratch J., Morency L.-P. Investigating Voice Quality as a Speaker-Independent Indicator of Depression and PTSD. Proceedings of Interspeech 2013. Lyon, 2013, p. 847–851.

[18] Kane J., Scherer S., Morency L.-P., Gobl C., A Comparative Study of Glottal Open Quotient Estimation Techniques. Proceedings of Interspeech 2013. Lyon, 2013, p. 1658–1662.

[19] Gong Y., Poellabauer C. Topic Modeling Based Multi-modal Depression Detection. Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC '17. Mountain View, 2017, pp. 69–76 DOI: 10.1145/3133944.3133945.

[20] Kane J., Gobl C. Identifying Regions of Non-Modal Phonation Using Features of the Wavelet Transform. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2011. Florence, 2011, pp. 177-180.

[21] Kane J., Gobl C. Wavelet Maxima Dispersion for Breathy to Tense Voice Discrimination. IEEE Transactions on Audio, Speech, and Language Processing, 2013, vol. 21, no. 6, pp. 1170–1179 DOI: 10.1109/TASL.2013.2245653.

[22] Cohn J. F., Kruez T. S., Matthews I., Yang Y., Nguyen M. H., Padilla M. T., Zhou F., De la Torre F. Detecting depression from facial actions and vocal prosody 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. Amsterdam, 2009, pp. 1–7, DOI: 10.1109/ACII.2009.5349358.

[23] Yu Z., Scherer S., Devault D., Gratch J., Stratou G., Morency L.-P., Cassell J. Multimodal Prediction of Psychological Disorders: Learning Verbal and Nonverbal Commonalities in Adjacency Pairs. Proceding of 17th Workshop Series on the Semantics and Pragmatics of Dialogue. Amsterdam, 2013.

[24] Lam G., Dongyan H., Lin W. Context-aware Deep Learning for Multi-modal Depression Detection. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, 2019, pp. 3946–3950, DOI: 10.1109/ICASSP.2019.8683027.

[25] Al Hanai T., Ghassemi M., Glass J. Detecting Depression with Audio/Text Sequence Modeling of Interviews. Proc. Interspeech 2018,

Hyderabad, 2018, pp. 1716–1720, DOI: 10.21437/Interspeech.2018-2522.

[26] Scherer S., Kane J., Gobl C., Schwenker F. Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. Computer Speech & Language, 2013, vol. 27, no. 1, pp. 263–287 DOI: 10.1016/j.csl.2012.06.001.

[27] Gratch J., Arstein R., Lucas G., Stratou G., Scherer S., Nazarian A., Wood R., Boberg J., DeVault D., Marsella S., Traum D., Rizzo A., Morency L. The Distress Analysis Interview Corpus of human and computer interviews. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, 2014, pp 3123–3128.

[28] DeVault D., Artstein R., Benn G., Dey T., Fast E., Gainer A., Georgila K., Gratch J., Hartholt A., Lhommet M., Lucas G., Marsella S., Morbini F., Nazarian A., Scherer S., Stratou G., Suri A., Traum D., Wood R., Xu Y., Rizzo A., Morency L. AAMAS '14: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems. Paris, 2014, pp. 1061–1068.

[29] SimSensei & MultiSense: Virtual Human and Multimodal Perception for Healthcare Support (2013). Available at: https://www.youtube.com/watch?v=ejczMs6b1Q4 (Accessed 30 November 2020).

[30] Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 2014, vol. 15, pp. 1929–1958.

[31] Keras documentation: Transfer learning & fine-tuning. (2020) Available at: https://keras.io/guides/transfer_learning/ (Accessed 30 November 2020).

[32]

Valverde-Albacete F. J., Peláez-Moreno C., 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox. PLoS One, 2014, vol. 9, no. 1, p. e84217 DOI: 10.1371/journal.pone.0084217.

[33] Lin T.-Y., Goyal P., Girshick R., He K., Dollár P. Focal Loss for Dense Object Detection. Computer Vision and Pattern Recognition, 2018. Available at: http://arxiv.org/abs/1708.02002. (Accessed 30 November 2020).

[34] tfa.losses.SigmoidFocalCrossEntropy | TensorFlow Addons TensorFlow (2020). Available at: https://www.tensorflow.org/addons/api_docs/python/tfa/losses/SigmoidFocalCrossEntropy (Accessed 30 November 2020).

[35] Mourão-Miranda J., Hardoon D. R., Hahn T., Marquand A. F., Williams S. C. R., Shawe-Taylor J., Brammer M. Patient classification as an outlier detection problem: An application of the One-Class Support Vector Machine. NeuroImage, 2011, vol. 58, no. 3, pp. 793–804, DOI: 10.1016/j.neuroimage.2011.06.042.

[36] Rondina J. M., Hahn T., de Oliveira L., Marquand A. F., Dresler T., Leitner T., Fallgatter A. J., Shawe-Taylor J., Mourao-Miranda J. SCoRS--A Method Based on Stability for Feature Selection and Mapping in Neuroimaging. IEEE Transactions on Medical Imaging, 2014, vol. 33, no. 1, pp. 85–98. DOI: 10.1109/TMI.2013.2281398.

[37] Hahn T., Marquand A. F., Ehlis A. C., Dresler T., Kittel-Schneider S., Jarczok T. A., Lesch K. P., Jakob P. M., Mourao-Miranda J., Brammer M. J., Fallgatter A. J. Integrating neurobiological markers of depression. Archives of general psychiatry, 2011, vol. 68, no. 4, pp. 361–368. DOI: 10.1001/archgenpsychiatry.2010.178.