An Unsupervised Learning Approach for Detecting Relapses from Spontaneous Speech in Patients with Psychosis

C. Garoufis[1], A. Zlatintsi[1], P. P. Filntisis[1], N. Efthymiou[1], E. Kalisperakis[2,3], V. Garyfalli[2,3], T. Karantinos[2], L. Mantonakis[2,3], N. Smyrnis[2,3] and P. Maragos[1]
[1]*School of ECE, National Technical University of Athens, 15773 Athens, Greece*
[2]*Laboratory of Cognitive Neuroscience, University Mental Health Research Institute, Athens, Greece*
[3]*National & Kapodistrian University of Athens, Medical School*
cgaroufis@gmail.com, {filby,nefthymiou}@central.ntua.gr, {nzlat, maragos}@cs.ntua.gr, smyrnis@med.uoa.gr

***Abstract*—In this work, we aim to explore and develop a speech analysis system that identifies relapses in patients with psychotic disorders (i.e., bipolar disorder and schizophrenia) with the long-term goal of monitoring and detecting relapse indicators, in order to aid in timely diagnoses of psychotic relapses. To this end, we utilize an unsupervised learning approach, employing convolutional autoencoders to build personalized speech models for patients. We use data from interviews between patients and clinicians to train and evaluate our models. The models are trained, learning to reconstruct spectrograms of speech segments corresponding to non-relapsing periods; then, the reconstruction error of the model is used to determine whether unseen speech data correspond to an anomalous (relapsing or pre-relapsing) state, or a stable one. A preliminary study using data from 5 patients and 95 interviews in total yielded encouraging results, indicating the potential usability of such models in real-time health monitoring.**

***Index Terms*—Psychotic Disorders, Mental Health, Anomaly Detection, Spontaneous Speech, Unsupervised Learning**

I. Introduction

During recent years, machine learning and artificial intelligence methods have been introduced into clinical psychiatry [1] and are increasingly being explored in various mental health contexts [2]. It is a fact that since no objective markers for psychotic disorder symptoms have yet been established, their severity assessments are mostly based on subjective reports requiring experienced clinicians [3], [4] and specialized questionnaires [5]. It is thus of critical importance to identify those markers (i.e., physiological, psychological, vocal, facial or social among others) that reflect mental health severity (eventually resulting in significant symptom burden and lower life expectancy [6]), in order to overcome these limitations.

In particular, speech has long been identified to contain patterns that largely diverge by a person's mood, leading to an increased focus on automated methods to monitor speech in patients with mental disorders [7]. For instance, characteristics such as hesitant, abrupt or low-voiced speech are considered aspects of psychomotor retardation, one of the prominent features of depression [8], [9]. Already during 1970s and 1980s, studies reported that depressed individuals showed such speech characteristics compared to healthy individuals [10], [11]; yet acoustic signal processing available at the time was not adequate to cross-validate the subjective observations [12].

In [13] various speech features that could characterize various psychotic illnesses are discussed; for instance, total time talking, speech rate, and mean pause duration measure poverty of speech and alogia and are considered as typical negative symptoms in schizophrenia. Flat affect, another negative symptom, could be expressed by lower f0 mean and variability. In bipolar disorder, significant increases in tonality were observed including increases in median f0 and mean F1 and F2, while a larger number of longer pauses was observed in depressive states than in euthymic or hypomanic states, with speech pauses becoming longer as patients enter depressive states. In other cases, pitch, formants, mel-frequency cepstrum coefficients (MFCC), linear prediction cepstral coefficient (LPCC), and gamma-tone frequency cepstral coefficients (GFCC), among others, have been extracted and classified using classic machine learning techniques, such as support vector machines (SVMs), for monitoring and predicting manic states in psychotic disorder patients [14].

As a result, during recent years, researchers using machine learning have succeeded in quantifying and modeling speech characteristics to assess the severity of depression [15], [16]. However, it is a challenging problem since characteristics like these may be influenced by the personality or speech habits of each subject, thus severely influencing the classifier performance. To overcome such limitations caused by handcrafted features, a variety of deep learning techniques have been proposed and employed for automatic feature extraction in order to model the evolution of emotional states from speech. To that end models such as CNNs, RNNs, autoencoders, as well as hybrid models have been employed to encode the temporal and spectral features from speech signals, to learn

This research has been financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH–CREATE–INNOVATE (project code:T1EDK-02890).

In 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, July 2021

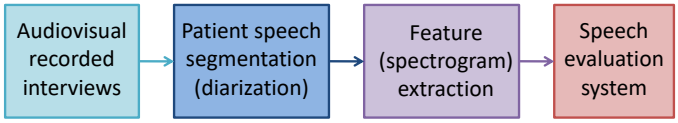

Fig. 1: An overview of the proposed speech analysis pipeline for relapse detection and prediction.

low-dimensional representations for vocal expressions, or to characterize the temporal evolution of emotions based on learned and/or other handcrafted features [17]–[19].

Another approach usually used in order to identify unusual measures (abnormalities, deviants, or outliers) in data is anomaly detection, used within diverse research areas and application domains [20]. The use of semi-supervised or unsupervised algorithms in anomaly detection is motivated by the scarcity of anomalous events as they are either very infrequent or completely unseen. Recent works have employed neural networks to detect anomalies in data with unknown distributions; in specific, autoencoders have been used in order to learn compressed representations of the data, with anomalies identified by higher reconstruction error [21]–[23].

In this work, we aim to explore and develop a speech analysis system that identifies pre-relapse and relapse states in patients with psychotic disorders (i.e., bipolar disorder and schizophrenia), in order to aid in effective monitoring and timely predictions and diagnoses of psychotic relapses. This work is part of e-Prevention, an ongoing research and development project[1]aiming to develop innovative and advanced e-Health services for medical support that will facilitate the effective treatment monitoring and the relapse prevention in such patients through processing i) biosignals from smartwatches that are acquired in a 24/7 basis, ii) audiovisual unstructured interviews (thus containing spontaneous speech) recorded weekly – through a dedicated tablet application – between patients and clinicians. Since research has shown that mood recognition can be improved by adapting a system to subject data, allowing it to pick up on subject-specific symptomatology [24], we opted for subject-dependent and personalized experiments. In our proposed approach, as presented in Fig. 1, features are extracted from patient utterances isolated from these interviews, and the utterances are then evaluated as clean or anomalous in a per-session basis.

We followed an unsupervised learning approach towards the evaluation of these utterances. In specific, we trained convolutional autoencoders in order to model the speech patterns of patients, using speech samples from interviews conducted in time periods where they had not been in a relapsing condition. Afterwards, we evaluated them in speech data from interview sessions taking place during both relapse and non-relapse periods, as well as a short time period before the occurrence of each relapse (i.e., pre-relapse). This has the additional advantage that, in contrast to supervised learning-based approaches, these models can be trained for each patient without the availability of speech data corresponding to relapse periods, which is usually the case during the first months of patient monitoring. A preliminary study of the system in 5 patients, having conducted a total of 95 such interview sessions, yields promising results, indicating that the above framework could not only be in assistance of the clinicians regarding the evaluation of the patients' condition, but potentially could also help in timely prediction of the relapses, thus reducing their severity.

[1]More info about the e-Prevention project: http://eprevention.gr

II. Data Collection

Twenty-four (24) patients with a disorder in the psychotic spectrum (12 with Schizophrenia, 8 with Bipolar I disorder, 2 with Schizoaffective disorder, 1 with Brief Psychotic episode, and 1 with Schizophreniform disorder) were recruited at the University Mental Health, Neurosciences and Precision Medicine Research Institute "Costas Stefanis" (UMHRI) in Athens, Greece. All volunteers, after being fully informed about the project's goals, signed a written consent for their participation and a written permission for the use of their personal (anonymized) data, in accordance with the provisions of the General Regulation (EU) 2016/679. Additionally, all protocols of the e-Prevention research project have been approved by the Ethics Committee of the Institution.

At recruitment, patients were in active treatment and stable and underwent an assessment of mental health symptoms and their general functioning. Afterwards and during the course of the project (since 28/05/2020), the clinicians have conducted with all patients a number of short (up to 10 min) "unstructured" interviews in a weekly or biweekly basis, in order to assess their physical activity by using the International Physical Activity Questionnaire – IPAQ-SF [25]. One of the innovative key offerings of the e-Prevention project regards the anonymous recording of these interviews through a tablet (a Samsung Galaxy Tab A6), installed in the patient's residence, with the goal to process this information and understand the emotional status of the patient. A dedicated application has been developed for this cause, in a way that would make it as simple as possible for the patients to use. After each interview, the recorded data are uploaded to a secure cloud infrastructure [26].

Additionally, the clinicians conducted in-person follow-up assessments with the patients once every month to administer various reliable rating scales regarding the general psychopathology (such as PANSS - Positive and Negative Syndrome Scale [27]) among others, so as to appraise the appearance and severity of psychotic relapses, when existing, providing this way valuable annotation data for the patient's mental health condition.

The annotations provided by the clinicians indicated the patient's condition as either stable or relapsing (also denoting the specific period of the relapse and its severity as low, mid or high). The determination of these relapses (i.e., the re-emergence of delusions, hallucinations in patients with a psy-

In 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, July 2021

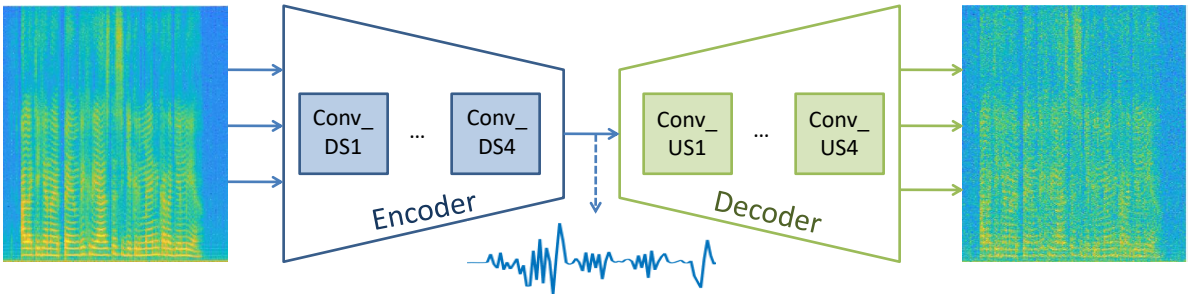

Fig. 2: Overview of the proposed autoencoder architecture, including the convolutional blocks inside the encoder and decoder.

TABLE I: Demographics information at the time of recruitment, and amount of recorded and analyzed data utterances.

Demographics	
Male/Female	2/3
Age (years)	30.8 ± 7.38
Education (years)	14.4 ± 1.88
Illness dur. (years)	7.2 ± 8.21
Recorded Data	
Num. of Interviews (total)	95
Num. of Interviews (mean±std)	19 ± 5.58
Diarized speech duration (in sec)	8795
Diarized sppech duration (in sec, mean±std)	1759.8 ± 1382.89
Num. of Utterances (total)	3467
Num. of Utterances (mean±std)	693.4 ± 421.7
Num. of Utterances (clean)	375 ± 261.3
Num. of Utterances (pre-relapse)	117.2 ± 91.5
Num. of Utterances (relapse)	201.2 ± 104

chotic disorder, the depression or the mania) was multifaceted; thus, it was taken into consideration and evaluated by the clinicians through the following: 1) The monthly assessments that assisted in quantifying the duration and severity of the relapse, and in determining the reason leading to it, 2) the use of PANSS that gives valuable information for the relapse itself, and communication with 3) the attending physician, 4) the family or the patient's carer and finally 5) with the hospital, in cases when the patient had to be hospitalized.

We have to note that since the interviews are unstructured, especially during a relapse, in this task we deal with spontaneous speech. Depending on the annotation data, these interviews were classified into three categories: clean data (C), where the condition of the patient was annotated as stable, relapse data (R), corresponding to the time periods where a relapse had occured and was thus annotated accordingly, and pre-relapse data (P), where the final category consists of the interviews that were held up to a month (specifically, 30 days) prior to the appearance of each relapse. For this study, we only used interview data from patients who had experienced a relapse during the study's duration – reducing the number of our subjects to five (5), including 1 with Schizoaffective disorder, 1 with Schizophreniform disorder, 1 with Schizophrenia and 2 with Bipolar I disorder. Table I contains information on the demographics of the patients, as well as the collected data at the time of writing this paper.

III. Methodology

A. Data Preprocessing and Feature Extraction

Firstly, the speech segments corresponding to each patient were isolated. To this end, we transformed the interview videos into .wav format, downsampled the audio to 16kHz, and performed speech diarization using the x-vector [28] diarization recipe from Kaldi [29], a widely-used speech processing toolkit, in order to isolate the utterances that belonged to the patients. After the diarization process, the extracted speech utterances were manually checked to ensure that the process was successful, and cleaned accordingly. The whole process resulted in a total of 3467 utterances, with a total length of 8795 seconds, spread between the five patients. The exact distribution of the extracted utterances in clean, relapsing and pre-relapsing states is presented in Table I.

Afterwards, using Librosa, we computed the mel-spectrograms over all the extracted utterances, with a frame size of 512 samples, a 256-sample overlap between successive windows, and 128 mel bands. Finally, the spectrograms were split into segments of 64 frames (coarsely equal to 1 sec), thus yielding an 128x64 representation for each speech segment.

B. Autoencoder Architecture

For each of the patients, in order to build a reliable speech model, we implemented a 2D-Convolutional Autoencoder, operating on the spectrograms derived as described in Sec. III-A. These networks consist of an encoder, that learns to map a high-dimensional input into a low-dimensional latent representation, and a decoder, which attempts to reconstruct the original input from the latent representation, as seen in Fig. 2.

In our case, the convolutional encoder consists of 4 successive downsampling convolutional blocks (denoted as Conv_DS), each of which includes a 2D-Convolution layer, a ReLU activation function, and a Max Pooling layer, to progressively reduce the dimensionality of the input. Similarly, the decoder consists of 4 successive upsampling convolutional blocks (denoted as Conv_US), which in turn incorporate Upsampling layers, in order to restore the initial dimensions of the input spectrogram, followed by 2D Convolution layers. ReLU activations were applied after each block as well, except for the final layer that uses a tanh activation function. Specific network parameters for each of the network blocks, including

In 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, July 2021

TABLE II: Architecture parameters for the convolutional autoencoder we employ in our study.

Net. Block	N	(k_x, k_y)	(p_x, p_y)	(u_x, u_y)	Output Dim.
Conv_DS1	128	(5,5)	(2,2)	-	64x32x128
Conv_DS2	256	(5,5)	(4,2)	-	16x16x256
Conv_DS3	512	(5,5)	(4,4)	-	4x4x512
Conv_DS4	1024	(5,5)	(4,4)	-	1x1x1024
Conv_US1	512	(5,5)	-	(4,4)	4x4x512
Conv_US2	256	(5,5)	-	(4,4)	16x16x256
Conv_US3	128	(5,5)	-	(4,2)	64x32x128
Conv_US4	1	(5,5)	-	(2,2)	128x64x1

the number of filters per convolutional layer, N, the kernel size of each convolution, (k_x, k_y), and the pooling (p_x, p_y) and upsampling (u_x, u_y) factors are presented in Table II.

C. Training and Evaluation Protocol and Metrics

A separate model was trained for each patient, using only speech segments corresponding to clean data. To assess the performance of our models, we need to compare the performance of each subject-dependent speech model between the speech segments corresponding to relapse, or pre-relapse periods, and "unseen" data annotated as clean. Thus, we performed K-fold cross-validation, using each time data from $K-1$ clean sessions to train the model, while evaluating the performance of the model on the spectrograms corresponding to the final clean session, as well as all spectrograms corresponding to sessions annotated as pre-relapsing or relapsing.

During each training fold, the training data were further split into a training and a validation set, on a per-session basis, using a $4:1$ session ratio. Min-max normalization was applied to the train set, so that the values of each spectrogram were scaled into a $[0, 1]$ range, and afterwards its parameters were applied to the spectrograms in the validation and testing sets.

The network models were implemented in Keras, trained using the Mean Square Error (MSE) as a loss function and the Adam optimizer [30] with a learning rate equal to 0.001. A batch size of 32 was used, while early stopping was applied to monitor the model's training, using the validation loss and patience of 10 epochs. As our primary evaluation metric, we use the average MSE between the reconstructed and the true spectrograms, aggregated over each session in the testing set. Smaller MSE values denote more accurate reconstruction, while larger ones indicate areas in the spectrogram that could not be correctly reconstructed, which could point to anomalies in the speech segments. Since each clean session is only treated as "unseen" data once, while every session corresponding to the period before or during a relapse is evaluated at each of the K folds, the MSEs of these sessions are further averaged over all K folds.

We finally note that the usage of the tanh activation function in the network's output allows the reconstructed spectrogram to take values in the $[-1, 1]$ range, instead of the allowed $[0, 1]$. Since the network is trained with $[0, 1]$-valued spectrograms, in most cases the values of the reconstructed spectrograms lie in the correct range. However some negative values still occur, and those are manually set equal to 0 when evaluating the model in data that were unseen during training.[2]

TABLE III: Per-session medians of the reconstruction mean-square error loss (MMSE) for each of the patients we studied and for a subject-independent model, depending on whether the state of the patient during the sessions is clean (C), pre-relapsing (P), or relapsing (R), as well as macro-F1 scores regarding the classification of the patient state as stable (clean) or anomalous.

Patient ID	MMSE (C)	MMSE (P)	MMSE (R)	macro-F1
#1	**0.00045**	0.00212	0.00197	0.71
#2	**0.00058**	0.00349	0.00073	0.61
#3	**0.00029**	0.00097	0.00309	0.80
#4	**0.00023**	0.00086	0.00042	0.73
#5	0.00537	**0.00270**	0.00420	0.60
Global	**0.00008**	**0.00008**	0.00010	0.61

IV. Results and Discussion

We present the results of our experiments in Table III. In specific, for each patient and each class (clean data, pre-relapse data and during-relapse data) we report on the median of the per-session aggregated reconstruction errors (MMSEs). We observe that for the majority of the patients involved, the median reconstruction error over sessions corresponding to the clean speech data is less than the respective error for the utterances corresponding to sessions where the patient is either in a relapsing state or a pre-relapsing one. We note that this trend does not hold for patient #5; we presume that this is due to the availability of less training data in comparison to the rest of the patients. We also note that the systems do not differentiate between segments during pre-relapsing or relapsing periods, however during both states the reconstruction errors compared to clean data are indeed higher.

In an attempt to quantify the margin of discriminability between the clean and anomalous data (corresponding to either a pre-relapsing or a relapsing state), we binarize the per-session mean square reconstruction errors using a suitable threshold value, and compare them to the ground truth labels of the sessions. These results are also reported in Table III, using the macro average F1-score as a metric. We observe that for all 5 patients, we achieve correct classification results in a percentage significantly higher than random chance, and specifically higher that 70% for three of the five patients, reaching up to 80% for one of the five patients. These results are comparable to those obtained by a number of studies on bipolar [7], [23] and depressive [19] patients, although we note that these studies followed a supervised learning protocol.

We further tested whether our proposed methodology can scale in a subject-independent setup; to this end, we trained a model on speech data corresponding to all patients (denoted as Global in Table III), and evaluated it using the same protocol

[2]Usage of the sigmoid activation function, which suppresses the spectrogram output values to $[0, 1]$, instead of the tanh led to occasional failure of the trained network to converge.

as above. On the one hand, lower reconstruction losses are recorded for this global model compared to the personalized ones, which was to be expected due to the larger amount of data used for training; however, the relapsing and pre-relapsing states are not as easily discriminated from the stable ones, offering an indication that relapse markers in speech appear in a subject-specific manner [24].

V. Conclusions

In this work, we presented a speech analysis system, utilizing unsupervised learning using convolutional autoencoders, in order to identify (anomalous) pre- and relapse states in patients with psychotic disorders (i.e., bipolar disorder and schizophrenia) with the ultimate goal of predicting relapse indicators, that could aid in timely diagnoses of psychotic relapses. We built speaker-dependent models to pick up on subject-specific symptomatology and our preliminary study with data from five patients that experienced a relapse during the course of the project yielded encouraging results regarding utterances recorded both during and prior to relapses, indicating that the approach we adopted is promising and could eventually assist in real-time health monitoring. In future analysis, we aim to explore other related databases in order to apply transfer learning techniques, and to experiment further with a variety of both unsupervised speech models (such as Variational or Recurrent Autoencoders) and input audio representations. In addition, we are interested in exploring multimodality in our data either using audio plus text (i.e., keywords), audio-visual information from the interviews, or even the combination of speech information with biosignals from smartwatches (signals that are also acquired 24/7 in the e-Prevention project).

References

[1] M. H. Aung, M. Matthews, and T. Choudhury, "Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies," *Depression and Anxiety*, vol. 34, no. 7, pp. 603–609, 2017.

[2] G.-D. Liu, Y.-C. Li, W. Zhang, and L. Zhang, "A Brief Review of Artificial Intelligence Applications and Algorithms for Psychiatric Disorders," *Engineering*, vol. 6, no. 4, pp. 462–467, 2020.

[3] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T.F. Quatieri, "A Review of Depression and Suicide Risk Assessment using Speech Analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[4] M. Yamamoto, A. Takamiya, K. Sawada, M. Yoshimura, Kitazawa, et al., "Using Speech Recognition Technology to Investigate the Association between Timing-related Speech Features and Depression Severity," *PloS one*, vol. 15, no. 9, pp. e0238726, 2020.

[5] M. Bauer, T. Wilson, K. Neuhaus, J. Sasse, Pfennig, et al., "Self-reporting software for bipolar disorder: validation of chronorecord by patients with mania," *Psychiatry research*, vol. 159, pp. 359–366, 2008.

[6] R. Emsley, B. Chiliza, L. Asmal, and B.H. Harvey, "The Nature of Relapse in Schizophrenia," *BMC Psychiatry*, vol. 13:1, pp. 1–8, 2013.

[7] J. Gideon, E. M. Provost, and M. McInnis, "Mood State Prediction from Speech of Varying Acoustic Quality for Individuals with Bipolar Disorder," in *Int'l Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Shanghai, China, 2016.

[8] D. Bennabi, P. Vandel, C. Papaxanthis, T. Pozzo, and E. Haffen, "Psychomotor Retardation in Depression: a Systematic Review of Diagnostic, Pathophysiologic, and Therapeutic Implications," *BioMed Research International*, vol. 2013, 2013.

[9] S. Scherer, G.M. Lucas, J. Gratch, A.S. Rizzo, and L.-P. Morency, "Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews," *IEEE Trans. on Affective Computing*, vol. 7, no. 1, pp. 59–73, 2015.

[10] E. Szabadi, C.M. Bradshaw, and J. Besson, "Elongation of Pause-Time in Speech: a Simple, Objective Measure of Motor Retardation in Depression," *The British Jour. of Psychiatry*, vol. 129, no. 6, pp. 592–597, 1976.

[11] R. Jouvent, D. Widlöcher, et al., "Speech Pause Time and the Retardation Rating Scale for Depression (ERD): Towards a Reciprocal Validation," *Jour. of Affective Disorders*, vol. 6, no. 1, pp. 123–127, 1984.

[12] J.C. Mundt, P.J. Snyder, M.S. Cannizzaro, K. Chappie, and D.S. Geralts, "Voice Acoustic Measures of Depression Severity and Treatment Response Collected via Interactive Voice Response (IVR) Technology," *Jour. of Neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[13] D.M. Low, K.H. Bentley, and S.S. Ghosh, "Automated Assessment of Psychiatric Disorders using Speech: A Systematic Review," *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[14] Z. Pan, C. Gui, J. Zhang, J. Zhu, and D. Cui, "Detecting Manic State of Bipolar Disorder Based on Support Vector Machine and Gaussian Mixture Model using Spontaneous Speech," *Psychiatry investigation*, vol. 15, no. 7, pp. 695, 2018.

[15] R.J. Davidson, D. Pizzagalli, J.B. Nitschke, and K. Putnam, "Depression: Perspectives from Affective Neuroscience," *Annual Review of Psychology*, vol. 53, no. 1, pp. 545–574, 2002.

[16] E. Goeleven, R. De Raedt, S. Baert, and E.H.W. Koster, "Deficient Inhibition of Emotional Information in Depression," *Journal of Affective Disorders*, vol. 93, no. 1-3, pp. 149–157, 2006.

[17] L. He and C. Cao, "Automated Depression Analysis using Convolutional Neural Networks from Speech," *Journal of Biomedical Informatics*, vol. 83, pp. 103–111, 2018.

[18] K.-Y. Huang, C.-H. Wu, and M.-H. Su, "Attention-Based Convolutional Neural Network and Long Short-Term Memory for Short-Term Detection of Mood Disorders Based on Elicited Speech Responses," *Pattern Recognition*, vol. 88, pp. 668–678, 2019.

[19] S. Harati, A. Crowell, H. Mayberg, and S. Nemati, "Depression Severity Classification from Speech Emotion," in *Proc. Int'l Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, USA, 2018.

[20] R. Chalapathy and S. Chawla, "Deep Learning for Anomaly Detection: A Survey," *arXiv preprint arXiv:1901.03407*, 2019.

[21] J. An and S. Cho, "Variational Autoencoder based Anomaly Detection using Reconstruction Probability," *Special Lecture on IE*, vol. 2:1, pp. 1–18, 2015.

[22] E. Rushe and B. Mac Namee, "Anomaly Detection in Raw Audio using Deep Autoregressive Networks," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[23] J. Gideon, K. Matton, S. Anderau, M.G. McInnis, and E.M. Provost, "When to Intervene: Detecting Abnormal Mood using Everyday Smartphone Conversations," *arXiv preprint arXiv:1909.11248*, 2019.

[24] S. Khorram, J. Gideon, M.G. McInnis, and E.M. Provost, "Recognition of Depression in Bipolar Disorder: Leveraging Cohort and Person-Specific Knowledge.," in *Proc. INTERSPEECH*, San Francisco, CA, USA, 2016.

[25] P. H. Lee, D. J. Macfarlane, T. H. Lam, and S. M. Stewart, "Validity of the intl. physical activity questionnaire short form (IPAQ-SF): A systematic review," *Int'l Jour. Behavioral Nutrition and Physical Activity*, vol. 8, pp. 115, 2011.

[26] I. Maglogiannis, A. Zlatintsi, A. Menychtas, D. Papadimatos, P. P. Filntisis, N. Efthymiou, G. Retsinas, P. Tsanakas, and P. Maragos, "An intelligent cloud-based platform for effective monitoring of patients with psychotic disorders," in *Proc. Int'l Conf. on Artificial Intelligence Applic. and Innovation*, Porto Carras, Greece, 2020.

[27] S.R. Kay, A. Fiszbein, and L.A. Opler, "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia," *Schizophrenia bulletin*, vol. 13, pp. 261–276, 1987.

[28] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AL, Canada, 2018.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, et al., "The Kaldi Speech Recognition Toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Waikiloa, HI, USA, 2011.

[30] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. Int'l Conf. on Learning Representations*, San Diego, CA, USA, 2015.

In 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, July 2021