Towards Unsupervised Subject-Independent Speech-Based Relapse Detection in Patients with Psychosis using Variational Autoencoders

C. Garoufis[1], A. Zlatintsi[1], P. P. Filntisis[1], N. Efthymiou[1], E. Kalisperakis[2,3], T. Karantinos[2], V. Garyfalli[2,3], M. Lazaridi[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@mail.ntua.gr, {filby,nefthymiou}@central.ntua.gr, {nzlat, maragos}@cs.ntua.gr, smyrnis@med.uoa.gr

***Abstract*—Generative models, such as Variational Autoencoders, are being increasingly utilized for various acoustic modeling tasks, such as anomaly detection from audio signals. Motivated by this, in this work we propose a Convolutional Variational Autoencoder (CVAE), in order to detect and predict the appearance of relapses in patients with psychotic disorders, such as schizophrenia and bipolar disorder. The proposed system utilizes speech segments of patients, isolated from interviews conducted with their clinicians containing spontaneous speech, and represented as log-mel spectrograms. The results from the analysis of each segment are then aggregated in a per-interview basis. We explore the performance of our system in both a personalized and a universal (patient-independent) setup. Evaluation of our method in data from 13 patients and 375 interviews, with a total duration of 30509 sec of isolated speech, indicate that the CVAE achieves similar results to a Convolutional Autoencoder (CAE) baseline in a personalized setup. Furthermore, the proposed model significantly outperforms the CAE baseline when considering a universal relapse detection setup.**

***Index Terms*—Psychotic Disorders, Anomaly Detection, Variational Autoencoder, Spontaneous Speech, Relapse Prediction**

I. Introduction

One of the many fields to have benefited from the rapid progress in the field of artificial intelligence is clinical psychiatry [1]. Machine learning algorithms are increasingly being applied in order to identify indicators of mental health severity, in order to validate the subjective assessments of clinicians [2], [3]. To this end, a number of modalities are being utilized, e.g., physiological data, facial expressions, and social activity information collected from smartphones.

Another modality that has been reported to contain cues correlated with the appearance of relapses in a number of mental conditions is speech. Indeed, spoken language has been shown to be indicative of both the emotional state of a person [4] and relapsing mental conditions [5], such as divergences in pitch, formant frequencies and pauses between utterances. For instance, bipolar disorder can be characterized by longer pauses in between utterances of the patients and increased pitch and formant frequencies, while relapsing schizophrenic patients show decreased pitch and formant frequencies, coupled with a lower speech rate and longer pauses in between utterances [5]. Supervised approaches towards determination of relapses from speech segments include both traditional hand-crafted features, either in a short-time basis [6] or aggregated over whole interview sessions [7] and deep-learning based approaches, usually utilizing convolutional or recurrent neural networks [8], [9]. The connection of the appearance of relapses in mental conditions with mood-related features through transfer learning, by employing pretrained emotionally imbued embeddings, has also been explored [10], [11].

An alternative approach to this problem, motivated by the rarity of appearance of abnormal (anomalous) events, as well the potential lack of strong labels concerns the development of either weakly-supervised or unsupervised anomaly detection algorithms [12]. This is especially significant in mental health monitoring, where the availability of data corresponding to relapsing states is scarce. Most recently deep neural networks have been used to detect anomalies in various kinds of medical settings, using for instance physiological signals [13], medical images [14] and speech signals [15], [16]. In the case of purely unsupervised algorithms, such as autoencoders, either properties of the learned latent vectors, or their reconstruction error, are usually used to evaluate their performance.

For the purposes of anomaly detection in audio signals, generative neural network architectures, such as the WaveNet [17], have been successfully adapted [18]. Variational Autoencoders (VAEs) [19], in particular, constitute a generative class of models that has proven suitable for various acoustic modeling tasks, as for instance speech enhancement [20], blind source separation [21], or speech representation learning [22]. It has also been leveraged in a mental health recognition context, using for instance phonocardiogram data [23].

In this work, we introduce a Convolutional VAE (CVAE) for the detection and prediction of psychotic patient relapses from spoken word segments. To the best of our knowledge, this is

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 acronym: e-Prevention, code:T1EDK-02890/ MIS: 5032797)

the first work that uses generative models, such as VAEs, in order to predict and detect relapses from speech. We explored both the performance of the CVAE in a *personalized* setup and its scalability to a *universal* (patient-independent) setting, by applying patient-wise normalization techniques similar to [7]. Experimental results indicate that in the case of *personalized* models, the CVAE-based models perform comparably to a deterministic CAE baseline [24]. In the *universal* case, the CVAE model significantly outperforms the CAE baseline, and surpasses the performance of the personalized models when patient-wise normalization is applied.

The rest of the paper is organized as follows: In Sec. 2 we describe in brief the database we use, and the preprocessing applied upon it, while the architecture of the network we developed is introduced in Sec. 3. We outline our experimental protocol in Sec. 4, and present and discuss our results and findings in Sec. 5. Finally in Sec. 6 we draw some final conclusions and present potential avenues for future research.

II. Data Collection and Preprocessing

Data Collection: A total of 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. The protocol regarding the recruitment of the patients in the project is detailed in [24].

During the course of the project, the clinicians have conducted monthly in-person clinical assessments with all patients. These assessments were used as the basis for the annotation of the patient's mental condition by the clinicians as either stable or relapsing. In particular, the appearance and severity of relapses were 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 usage of psychopathological scales that provide valuable information for the relapse itself, and communication with 3) the attending physician, 4) the family or the patient's carer and 5) with the hospital, upon the patient's hospitalization. In addition, a number of weekly unstructured interviews, of an average duration of 5-10 minutes, was conducted with all patients, These interviews were recorded anonymously through a dedicated tablet application, and stored into a secure cloud server [25].

Our goal in this work is to identify and predict the appearance of relapses in these patients. To evaluate this, we used interview data from 13 patients (1 with Schizoaffective disorder, 1 with Schizophreniform disorder, 7 with Schizophrenia and 4 with Bipolar I disorder), out of whom 8 had experienced a relapse during the course of the study, and the rest were selected on the basis of the available data amount. Table I contains information on the patient demographics and the collected data at the time of writing this paper. The expert annotations were used as the basis for splitting the data. In particular, we split the interviews into three categories: **clean** data, which correspond to time periods the patient's condition was stable, **relapse** data, where a relapse has been detected by the clinicians, and **pre-relapse** data, which include interviews conducted up to 30 days prior to the appearance of a relapse.

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

Demographics	
Male/Female	8/5
Age (years)	27.5 ± 6.7
Education (years)	13.5 ± 1.9
Illness dur. (years)	7.9 ± 7.6
Recorded Data	
Num. of Interviews (total)	375
Num. of Interviews (mean±std)	28.8 ± 8.7
Diarized speech duration (in sec)	30509
Diarized speech duration (in sec, mean±std)	2347 ± 1550
Num. of Utterances (total)	12107
Num. of Utterances (mean±std)	931 ± 527
Num. of Utterances (clean, mean±std)	754 ± 425
Num. of Utterances (pre-relapse, mean±std)	119 ± 126
Num. of Utterances (relapse, mean±std)	169 ± 162

Preprocessing & Feature Extraction: Regarding the preprocessing of the interviews, in order to extract meaningful representations of the patients' speech, we follow the same procedure as in [24]. In brief, we extracted the audio from the interview videos, and downsampled it to 16 kHz. Afterwards, the speech segments corresponding to the patients were isolated using the x-vector [26] diarization recipe from kaldi [27], and were then manually checked for correctness. This process resulted in a total of 12107 utterances (30509 sec), distributed into each category as presented in Table I.

Afterwards, we computed the log-scaled mel-spectrograms corresponding to each utterance, using Librosa [28]. For this computation, we used a frame size of 512 samples (approx. 30 ms), an overlap of 256 samples, and 128 mel bands. Finally, the per-utterance spectrograms were cut along the temporal axis into slices of 64 frames (approx. 1 sec), resulting in an $128x64$ representation for each 1 sec. of speech.

III. Network Architecture

A. Variational Autoencoders

Variational Autoencoders (VAEs) are a family of autoencoder architectures first developed in [19], and can be viewed as a probablistic variant of the classical autoencoders. Similarly to those, they consist of two parts, an encoder and a decoder. In more detail, the encoder (inference model) takes as input a tensor X, and encodes it into a latent representation $z \in \mathbb{R}^D$, that is assumed to follow an isotropic Gaussian distribution, conditional on the input tensor X:

$$q_\phi(z_i|X) = N(\mu_i, \sigma_i^2), i = 1, ..., D, \quad (1)$$

where the parameters μ_i, σ_i are learned via backpropagation and ϕ denotes the posterior parameters of the encoder. The decoder (generative model), on the other hand, samples an element from the above distribution, and attempts to learn a set of parameters θ, so that the marginal likelihood $p_\theta(X)$ is maximized. The marginal likelihood can be factorized as:

$$p_\theta(X) = p_\theta(z)p_\theta(X|z), \quad (2)$$

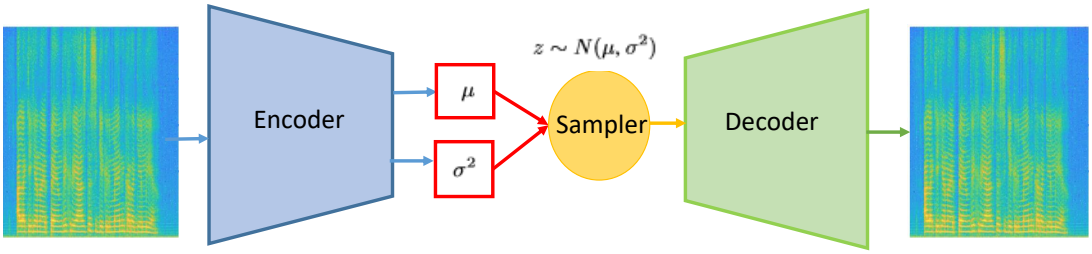

Fig. 1. An overview of the proposed variational autoencoder architecture.

where $p_\theta(z)$ corresponds to the prior distribution of the latent variable z. Thus, maximizing the marginal (log-)likelihood equates to maximizing the quantity:

$$\text{argmax}_{\theta,\phi}[E_{q(z|X)}(\ln(p_\theta(X|z)) - D_{KL}(q_\phi(z|X), p_\theta(z))] \quad (3)$$

Assuming a spherical isotropic Gaussian prior, $p_\theta(z) = N(0, I)$, for the latent vector z, maximizing the first term equates to minimizing the mean square error (MSE) between the true data X and the observed through inference data $\hat{X}$:

$$\mathcal{L}_{MSE} = ||\hat{X} - X||^2, \quad (4)$$

while maximizing the second term of (3) minimizes the Kullback-Leibler (KL) divergence between the learned posterior $q_\phi(z|X)$ and the assumed prior $p_\theta(z)$ of the latent vector.

B. Architecture Details

The anomaly detection model used in this paper is a 2D Convolutional Variational Autoencoder (CVAE), trained to reconstruct input spectrograms. An overview of the architecture is given in Table II. The encoder receives as input log-mel spectrograms, and consists of 3 downsampling convolutional blocks, which alternately extract intermediate features from the input spectrograms using Convolutional layers with ReLUs as the respective activation function and reduce the resolution of these feature maps with Max Pooling layers. These are followed by two parallel layers that estimate the parameters (μ, σ) of the Gaussian distribution of the latent vector z.

The decoder samples a latent vector $\hat{z}$ from the learned latent distribution through the reparameterization trick, to enable backward propagation of the gradients during training:

$$\hat{z} = \mu + \epsilon\sigma^2, \epsilon \sim N(0, I). \quad (5)$$

This latent vector is then propagated forward through 4 upsampling convolutional blocks (Conv_US in Table II), each of which upsamples the feature map it receives, and then processes the upsampled feature map through a 2D-Convolutional layer. ReLU activation functions are applied after each Convolutional layer, with the exception of the last one where no activation function is applied.

IV. Experimental Protocol

We train models for both the *personalized* case, where a unique model is trained for each patient, and the *universal* case, where a single model is trained using data from all patients. We follow a 5-fold cross-validation protocol, with data from segments both during and prior to the appearance of a relapse being considered as anomalous. In particular, for each fold, the data corresponding to stable time periods were split into training, validation and testing data by a 3:1:1 ratio, and the test set was afterwards injected with the data corresponding to both relapses and pre-relapse periods. We further note that the data were split so that spectrograms corresponding to the same interview session belonged in the same fold. In contrast to [24], we noted that feature-wise min-max normalization led the model to collapse its posterior into the uninformative isotropic Gaussian prior – thus, feature-wise standard scaling was applied to the log spectrograms. All models were trained using Keras, using Adam with a learning rate equal to 0.0003, and a batch size of 8. Training took place for a maximum of 200 epochs, with early stopping applied after 10 epochs with no improvement on the validation loss. After preliminary experiments, the loss weights corresponding to the MSE loss and the KL divergence were set to $\alpha = 1$ and $\beta = 0.01$.

TABLE II
Architecture parameters of the CVAE, including the number of filters, N_{filt}, the kernel sizes, (k_x, k_y), and the pooling (p_x, p_y) or upsampling (u_x, u_y) factors for each layer.

Net. Block	N_{filt}	(k_x, k_y)	(p_x, p_y)	(u_x, u_y)
Conv_DS1	N	(5,5)	(2,2)	-
Conv_DS2	2N	(5,5)	(4,2)	-
Conv_DS3	4N	(5,5)	(4,4)	-
Conv_DS4(μ)	8N	(4,4)	(4,4)	-
Conv_DS4(σ)	8N	(4,4)	(4,4)	-
Sample	-	-	-	-
Conv_US1	4N	(5,5)	-	(4,4)
Conv_US2	2N	(5,5)	-	(4,4)
Conv_US3	N	(5,5)	-	(4,2)
Conv_US4	1	(5,5)	-	(2,2)

We compare the performance of the CVAE to the deterministic CAE presented in [24], using the same hyperparameters for both networks ($N = 32$ filters and standardized log-scaled mel spectrograms). During evaluation, each element of the testing set is assigned an anomaly score. In the case of the CAE, the anomaly score is derived from the mean square reconstruction error (MSE) of the spectrogram. On the other hand, the performance of the CVAE is evaluated by either the KL divergence between the distribution of spectrogram's latent representation and its assumed prior, or its reconstruction MSE. To characterize each session as clean or anomalous, we aggregate the anomaly scores over all spectrograms corresponding to this session. As our evaluation metric, we report on the mean ROC-AUC over the per-session anomaly scores, since it is independent of threshold values [29].

TABLE III
AVERAGE OF THE PER-PATIENT ROC AUC SCORES FOR THE DISCRIMINATION BETWEEN SESSIONS THAT CORRESPOND TO STABLE, OR ANOMALOUS (PRE-RELAPSING OR RELAPSING) CONDITION, FOR BOTH CVAE AND CAE PERSONALIZED MODELS.

Pooling Function	CAE [24]	CVAE	
		MSE	KL
AP	**0.668** $\pm$ 0.035	**0.673** $\pm$ 0.055	0.653 $\pm$ 0.052
MP	0.608 $\pm$ 0.060	0.617 $\pm$ 0.051	0.659 $\pm$ 0.045
NP	0.627 $\pm$ 0.058	0.640 $\pm$ 0.049	**0.678** $\pm$ 0.049

In the subsequent experiments, we want to examine the following:

- How does the CVAE model compare to the deterministic AE used in [24], in both personalized and universal cases?
- In the universal case, how is the performance affected by whether the spectrogram normalization is applied globally (using the same transform parameters for all patients) or in a per-patient basis (computing separate transform parameters for each patient)?
- What type of temporal pooling is the most suitable for aggregating the anomaly scores of each spectrogram belonging to a specific session? With regard to this point, we examine average pooling (AP), max pooling (MP) and a non-learned variant of norm pooling (NP) [30], behaving similarly to softmax pooling and defined as:

$$S = \frac{(\sum_{i=1}^{N} |x^p[i]|)^{1/p}}{N} \tag{6}$$

where $\mathbf{x}$ is defined as the vector of per-spectrogram anomaly scores, and p is a positive integer. After preliminary experiments, we used the value $p = 10$.

V. RESULTS AND DISCUSSION

Personalized Models: In the case of personalized models, we train a unique model for each patient that has experienced a relapse during the course of the study. In Table III, we present the macro-average of the per-patient ROC-AUC scores of the baseline model (CAE) [24], as well as the proposed CVAE, and all three potential temporal pooling functions. The results indicate that in the personalized case, the CVAE-based models perform equally well to the original CAE, with no statistically signifcant difference ($p > 0.05$) found between the tested models. In addition, while in both the CAE and MSE-based CVAE models the average temporal pooling outperforms both max pooling and norm pooling-based variants, when using the KL divergence as an anomaly measure, we obtain better results when using the norm pooling. The per-patient ROC-AUC scores, using the best performing temporal aggregation function for each model, are presented in Table IV. We observe that for 6 out of the 8 patients, the personalized CVAE models based on the KL-divergence record a ROC-AUC score above 0.65, and for 3 out of the 8 patients, above 0.75.

Universal Models: In the case of the universal models, we train one model using spectrograms from all patients, regardless of whether they have experienced a relapse. In Table V, we present the ROC-AUC scores depending on i) whether the normalization procedure was applied globally or in a personalized manner, and ii) the temporal pooling function used to aggregate the per-spectrogram anomaly scores.

TABLE IV
PER-PATIENT ROC AUC SCORES FOR THE DISCRIMINATION BETWEEN SESSIONS THAT CORRESPOND TO STABLE, OR ANOMALOUS, CONDITION, FOR BOTH CVAE AND CAE PERSONALIZED MODELS.

Patient ID	CAE [24]	CVAE	
		MSE	KL
#1	0.546 $\pm$ 0.069	**0.547** $\pm$ 0.081	0.523 $\pm$ 0.134
#2	**0.448** $\pm$ 0.093	0.418 $\pm$ 0.182	0.388 $\pm$ 0.120
#3	**0.711** $\pm$ 0.119	0.700 $\pm$ 0.159	0.656 $\pm$ 0.187
#4	0.676 $\pm$ 0.066	0.660 $\pm$ 0.034	**0.768** $\pm$ 0.066
#5	0.781 $\pm$ 0.053	**0.800** $\pm$ 0.095	0.774 $\pm$ 0.040
#6	0.512 $\pm$ 0.067	0.520 $\pm$ 0.173	**0.696** $\pm$ 0.083
#7	0.877 $\pm$ 0.076	**0.940** $\pm$ 0.074	0.720 $\pm$ 0.186
#8	0.800 $\pm$ 0.187	0.850 $\pm$ 0.292	**0.900** $\pm$ 0.200
Average	0.668 $\pm$ 0.035	0.673 $\pm$ 0.055	**0.678** $\pm$ 0.049

Upon inspection of the results, we can observe that in this case, the CVAE consistently outperforms the CAE when the KL divergence is used to compute the anomaly score, irrespective of the temporal pooling function used. Application of the paired t-test between the CAE baseline and the two CVAE-variants showed statistically significant ($p < 0.05$) difference in the performance of the baseline and KL-divergence based CVAE model when per-patient normalization was applied to the input spectrograms. Moreover, in this case, the performance of the universal CVAE is at least comparable to the personalized models. The usage of the average pooling function as an aggregator yields the worst results, while the best results are given when using norm pooling with the KL-divergence scores, approaching a ROC-AUC score of 0.7. The superior performance of the CVAE in the universal setup indicates that the CVAE is able to learn subject-independent features at its bottleneck, a finding that is in agreement with VAEs being able to extract speaker-invariant representations from speech signals [22]. On the other hand, the dependence of both models to per-patient normalization indicates a limitation regarding their performance in a setting with unseen subjects.

Qualitiative Analysis: In Fig. 2, we display the log-scaled per-spectrogram KL loss for two session excerpts of the same patient, one corresponding to stable (dashed blue) and one to relapsing (orange) condition, as estimated by a universal model. We observe that the relapsing session does not record consistently higher anomaly scores, with the exception of a few peaks, highlighted in red. Upon observation of the respective audio segments, we notice that these correspond to temporary disruptions of the patient's speech flow, during the same utterance.

VI. CONCLUSIONS

In this work, we explored the potential of CVAEs, a class of generative models, in detecting and predicting relapses in patients with psychotic disorders from spontaneous speech. The results indicate that in the personalized case, these models work equivalently to a deterministic CAE baseline. In the universal case, they achieve at least comparable performance to the personalized ones when a per-patient normalization

TABLE V
ROC AUC SCORES FOR THE DISCRIMINATION BETWEEN SESSIONS THAT CORRESPOND TO STABLE, OR ANOMALOUS, CONDITION, FOR BOTH CVAE AND CAE UNIVERSAL MODELS, DEPENDING ON THE NORMALIZATION PROTOCOL AND POOLING FUNCTION USED.

Pers. Norm	Pool. Func.	CAE [24]	CVAE	
			MSE	KL
✗	AP	0.504 ± 0.032	0.502 ± 0.016	**0.532** ± 0.023
✗	MP	0.542 ± 0.024	0.551 ± 0.023	**0.592** ± 0.031
✗	NP	0.531 ± 0.034	0.527 ± 0.024	**0.581** ± 0.036
✓	AP	0.552 ± 0.036	0.585 ± 0.021	**0.646** ± 0.035
✓	MP	0.541 ± 0.027	0.622 ± 0.034	**0.685** ± 0.040
✓	NP	0.542 ± 0.028	0.618 ± 0.030	**0.698** ± 0.042

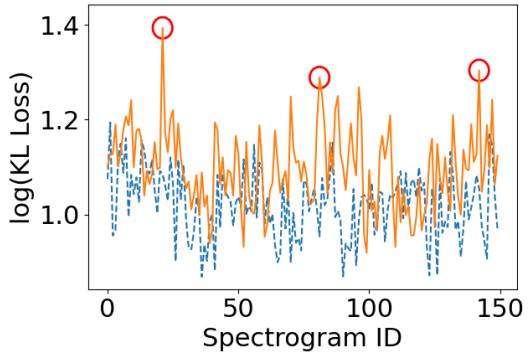

Fig. 2. Per-spectrogram visualization of the KL divergence scores for two sessions of the same patient, one corresponding to stable (dashed blue) and one to relapsing (orange) condition.

protocol was followed, significantly outperforming the CAE baseline, while using norm pooling to aggregate the per-session results further improves performance. Future work could focus into utilizing multimodal information for the detection and prediction of relapses, such as text transcripts of the interviews or biosignals, or taking advantage of longer-term dependencies in the interviews.

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

[3] M. Yamamoto *et al.*, "Using Speech Recognition Technology to Investigate the Association between Timing-related Speech Features and Depression Severity," *PloS one*, vol. 15, no. 9, p. e0238726, 2020.

[4] E. Szabadi, C. 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.

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

[6] 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, p. 695, 2018.

[7] J. Gideon, E. M. Provost, and M. McInnis, "Mood State Prediction from Speech of Varying Acoustic Quality for Individuals with Bipolar Disorder," in *Proc. ICASSP 2016*, Shanghai, China, 2016.

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

[9] 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.

[10] 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.

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

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

[13] K. Wang et al., "Research on Healthy Anomaly Detection Model Based on Deep Learning from Multiple Time-Series Physiological Signals," *Scientific Programm.*, 2016.

[14] A. Esteva et al., "Dermatologist-level Classification of Skin Cancer with Deep Neural Networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

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

[16] J. Vasquez-Correa, T. Arias-Vergara, J. O.-A. M. Schuster, and E. Nöth, "Parallel Representation Learning for the Classification of Pathological Speech: Studies on Parkinson's Disease and Cleft Lip and Palate," *Speech Communication*, vol. 122, pp. 56–67, 2020.

[17] A. van den Oord et al, "Wavenet: A Generative Model for Raw Audio," *arXiv preprint arXiv:1609.03499*, 2016.

[18] E. Rushe and B. Mac Namee, "Anomaly Detection in Raw Audio using Deep Autoregressive Networks," in *Proc. ICASSP 2019*, Brighton, UK, 2019.

[19] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[20] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational Autoencoder for Speech Enhancement with a Noise-Aware Encoder," in *Proc. ICASSP 2021*, Montreal, QU, Canada, 2021.

[21] J. Neri, R. Badeau, and P. Depalle, "Unsupervised Blind Source Separation with Variational Auto-Encoders," in *Proc. EUSIPCO 2021*, Dublin, Ireland, 2021.

[22] J. Chorowski, R. Weiss, S. Bengio, and A. van den Oord, "Unsupervised Speech Representation Learning using WaveNet Autoencoders," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 27, no. 12, pp. 2041–2053, 2019.

[23] R. Banerjee and A. Ghose, "A semi-supervised Approach for Identifying Abnormal Heart Sounds using Variational Autoencoder," in *Proc. ICASSP 2020*, Barcelona, Spain, 2020.

[24] C. Garoufis, A. Zlatnitsi, P. Filntisis, N. Efthymiou, E. Kalisperakis, T. Karantinos, V. Garyfalli, L. Mantonakis, N. Smyrnis, and P. Maragos, "An Unsupervised Learning Approach for Detecting Relapses from Spontaneous Speech in Patients with Psychosis," in *Proc. BHI 2021*, Athens, Greece, 2021.

[25] I. Maglogiannis et al., "An Intelligent Cloud-Based Platform for Effective Monitoring of Patients with Psychotic Disorders," in *Proc. Int. Conf. on Artif. Intell. Applic. and Innovation*, Porto Carras, Greece, 2020.

[26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proc. ICASSP 2018*, Calgary, AL, Canada, 2018.

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

[28] B. McFee *et al.*, "Librosa: Audio and Nusic Signal Analysis in Python," in *Proceedings of the 14th Python in Science Conference*, Austin, TX, USA, 2015.

[29] D. Fourure, M. Javaid, N. Posocco, and S. Tihon, "Anomaly Detection: How to Artificially Increase your F1-Score with a Biased Evaluation Protocol," in *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*, Bilbao, Spain, 2021.

[30] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio, "Learned-norm Pooling for Deep Feedforward and Recurrent Neural Networks," in *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*, Nancy, France, 2014.

[31] M. Rezaei, H. Yang, and C. Meinel, "Conditional Generative Refinement Adversarial Networks for Unbalanced Medical Image Semantic Segmentation," *arXiv preprint arXiv:1810.03871*, 2018.