Feature-Level Multimodal Fusion for Relapse Detection in Patients with Psychosis

Artemis Androni[2], Christos Garoufis[1,2,3], Athanasia Zlatintsi[1,2,3], Petros Maragos[1,2,3]
[1]*Robotics Institute, Athena Research Center, Athens, Greece*
[2]*School of ECE, National Technical University of Athens, Athens, Greece*
[3]*HERON - Center of Excellence in Robotics, Athens, Greece*
artemisandroni@gmail.com, cgaroufis@mail.ntua.gr, {nzlat, maragos}@cs.ntua.gr

***Abstract*—In this work, we aim to improve the detection of relapses in patients with psychotic disorders (i.e., bipolar disorder and schizophrenia) using spontaneous speech data from patient–clinician interviews and physiological signals from wearable sensors. To achieve this, we propose a joint autoencoder framework, trained with coupled data from both modalities, projecting them through distinct encoder branches into a unified latent space and then separately decoding them. We experiment with convolutional and LSTM-based autoencoders for the speech data, whereas we adapt convolutional autoencoders for the physiological signals. Our experimental results show that the proposed multimodal fusion scheme consistently outperforms unimodal baselines, with the introduced LSTM-based autoencoders proving to be an effective alternative to convolutional ones for relapse detection from speech. Moreover, through ablation experiments, we confirm that not only both modalities contribute positively to the joint framework, but our approach outperforms unimodally-trained baselines when only the respective modality is available.**

I. Introduction

Advancements in machine learning and artificial intelligence in recent years have transformed healthcare, including the field of clinical psychiatry [1]. Traditional diagnostic methods based on clinical evaluations, self-reports, and questionnaires [2] remain essential, yet they can miss the subtle early indicators of relapse in severe mental health disorders. By leveraging information from vocal patterns, physiological signals, and behavioral cues, modern machine learning approaches have been able to detect subtle signs of relapse, thereby bolstering conventional assessments [3], [4], and leading to improved patient outcomes.

Digital phenotyping, i.e., the collection and identification of physiological or behavioral markers from smart devices [5], has emerged as a promising approach for relapse detection. Supervised approaches, using algorithms such as Naive Bayes, k-Nearest Neighbors, and XGBoost [6], have been applied to smartphone sensor data to predict state transitions in bipolar disorder and depression [7], [8]. Additionally, autoencoder-based models have been used to identify deviations in such markers as relapse indicators [9], [10].

Regarding speech-based analysis, acoustic features such as pitch, MFCCs, and LPCCs have been extracted from spontaneous speech to detect manic states in bipolar disorder via Support Vector Machines (SVMs) and Gaussian Mixture Models (GMMs) [11]. Furthermore, deep learning architectures, such as convolutional and Long Short-Term Memory (LSTM) neural networks, have been employed for relapse detection in psychotic patients [4], [12], as well as for capturing emotional cues from speech and measure the severity of depression, achieving high accuracy in mood disorder detection [13], [14].

Mental health conditions are multifaceted with relapses manifesting across behavioral, physiological, and vocal domains [15], [16], highlighting the need for multimodal fusion. Several approaches have combined audio, visual, and textual features to successfully enhance depression detection using various deep learning architectures [17], [18], [19], [20]. Moreover, the fusion of textual, behavioral, and visual data from online social networks has demonstrated improved performance over unimodal approaches in depression detection [21].

The work presented in this paper builds upon the research conducted during the course of the e-Prevention project [22], where long-term biometric data from wearable sensors and audio-visual data from clinical interviews were continuously collected and analyzed to identify relapse-related markers, aiming to enable effective monitoring and relapse prevention for patients in the psychotic spectrum. As a part of the project, anomaly detection algorithms, operating in either spontaneous speech signals from patient-clinician interviews [12] or physiological signals derived from smartwatch sensors (i.e., accelerometer, gyroscope, and heart rate) [10], were developed for the detection of relapsing states. In this work, we combine these modalities in a unified, end-to-end trainable framework, to improve the relapse detection of these states. In more detail, our main contributions are as follows:

- Inspired by [23], we develop LSTM-based autoencoders, alongside the previously used convolutional autoencoders for relapse detection in speech data [12], benchmarking their performance in the – expanded with new patient data and relapse cases – audio portion of the e-Prevention database [22].
- We develop joint autoencoder models, consisting of an audio branch and a physiological branch, that perform feature-level fusion of the two modalities, deviating from the late-fusion setup presented in [22].

Experimental results indicate that our proposed feature-level fusion approach offers richer feature representations than unimodal models, leading to enhanced relapse detection. Furthermore, experiments in which individual branches

979-8-3315-1213-2/25/$31.00 ©2025 IEEE

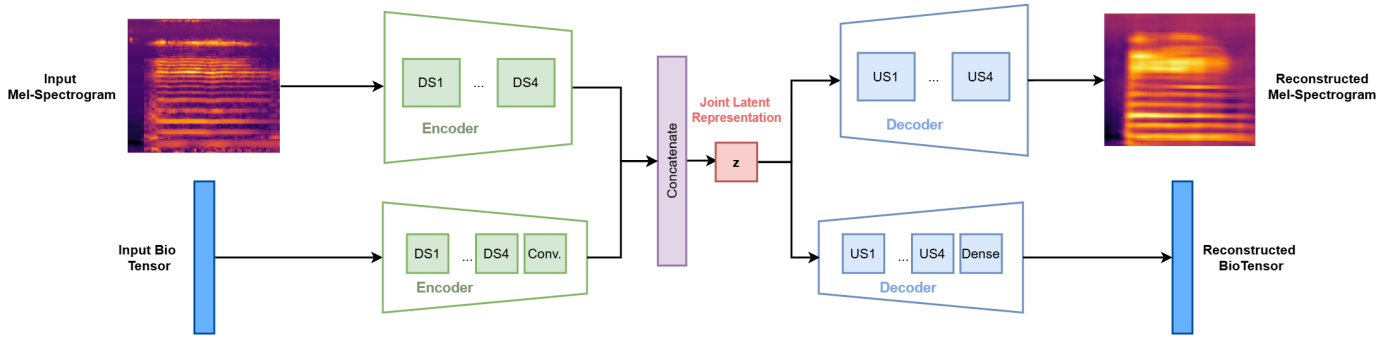

Fig. 1. Overview of the proposed multimodal autoencoder framework, for the case of a convolutional audio backbone; speech spectrograms and physiological data are separately encoded into a common low-dimensional latent space, and are subsequently reconstructed through the respective decoders.

were disabled demonstrated the essential contributions of both modalities, as well as its robustness to missing modality cases. Finally, LSTM-based autoencoders proved to be an effective alternative approach for detecting relapse from speech data.

II. Data Collection & Preprocessing

A. Data Collection

A total of thirty-nine (39) patients diagnosed with psychotic spectrum disorders, i.e., schizophrenia and bipolar disorder. were recruited for the purposes of the e-Prevention study; the recruitment protocol is delineated in [22]. For biometric data collection, participants wore Samsung Gear S3 Frontier smartwatches, which continuously recorded physiological and behavioral data, including heart rate and movement activity during both wakefulness and sleep. Additionally, for a subset of the participants, weekly or biweekly unstructured interviews were conducted between these participants and the clinicians. The interviews, averaging 5–10 minutes each, were conducted via a dedicated tablet application or telephone to assess physical activity using the Greek version of the International Physical Activity Questionnaire (IPAQ-Gr) [24]; video recordings of these interviews were anonymized and securely stored on a cloud server [25]. Moreover, clinicians conducted monthly in-person follow-up assessments using established rating scales, such as the Positive and Negative Syndrome Scale (PANSS) [26], to evaluate overall psychopathology and quantify the presence and severity of psychotic relapses, thereby generating annotations reflecting each patient's mental health state. Detailed criteria for these annotations are provided in [22]. Based on the clinicians' annotations, each patient's data were split into three categories: **clean** data, representing stable conditions, **relapse** data, corresponding to periods with annotated relapses and **pre-relapse** data, recorded up to 28 days before a confirmed relapse. In the context of this work, both relapse and pre-relapse states are considered anomalous.

In our work, the objective is to detect the appearance of relapses in patients from the e-Prevention database using these annotations as ground truth. For the audio-only autoencoder models, after expanding the e-Prevention database [22], we used data from 9 patients having experienced a relapse during the course of the project and whose demographics are presented in Table I; total data used correspond to 192 clean, 27 pre-relapse, and 42 relapse sessions. For the multimodal experiments, and taking into account the availability of adequate physiological data, we used data from 7 out of 9 patients, amounting to 3,280 hours of recorded physiological data.

TABLE I. Demographics and illness information for relapse patients in the extended e-Prevention database.

Demographics	
Male/Female	4/5
Age (years)	28.1 ± 7.6
Education (years)	13.3 ± 1.9
Illness duration (years)	6.6 ± 7.2

B. Data Preprocessing

Audio Data: Audio was extracted from patient–clinician interview videos and downsampled to 16 kHz to standardize the recordings. To isolate patient speech from that of the clinicians, we applied the x-vector [27] diarization pipeline from the Kaldi toolkit [28] and manually reviewed and corrected segments, where necessary. Next, the isolated speech segments were processed using Librosa to compute log-mel spectrograms with a frame length of 512 samples, a hop length of 256 samples, and 128 mel bands. Finally, each spectrogram was divided into fixed-length segments of 64 frames, resulting in a 128×64 feature representation for every second of speech.

Physiological Data: The accelerometer and gyroscope data, sampled at 20 Hz, and the heart rate sensor, sampled at 5 Hz, were first reviewed to ensure adequate data availability – discarding any days with less than four hours of recordings. To optimize the use of available data, each day of recorded data was divided into 8-hour segments, retaining only those segments containing at least four hours of valid data. Regarding feature extraction, a set of 10 features was computed within 5-minute windows. These features include the short-time energy of the accelerometer and gyroscope signals, the mean heart rate and R-R interval, as well as the power ratios of the low-frequency (0.04–0.15 Hz) and high-frequency (0.15–0.4 Hz) bands of the Lomb–Scargle periodogram [29]. Additionally, we extracted the width of the ellipse from the Poincaré recurrence plot, the ratio of valid samples within each 5-minute interval, and the sinusoidal representation of the corresponding seconds to capture chronological patterns. This process resulted in a feature tensor of size 96×10 for each 8-hour segment.

III. Methodology

We present an overview of our proposed framework in Fig. 1. In more detail, it consists of two separate encoder branches, one processing the mel-spectrograms and one the physiological feature tensors, in order to generate latent representations. These latent feature representations are then concatenated into a unified latent space, which is subsequently

TABLE II. COMPARISON OF ROC-AUC SCORES FOR THE PERSONALIZED CONVOLUTIONAL AUTOENCODER (CAE) AND LSTM AUTOENCODER (LSTMAE).

Patient ID	ROC-AUC	
	CAE	LSTMAE
#1	0.653 ± 0.048	**0.668** ± 0.050
#2	0.468 ± 0.159	**0.500** ± 0.087
#3	0.722 ± 0.165	**0.733** ± 0.231
#4	0.650 ± 0.093	**0.653** ± 0.135
#5	**0.754** ± 0.082	0.727 ± 0.081
#6	0.500 ± 0.159	**0.510** ± 0.183
#7	0.905 ± 0.055	**0.958** ± 0.093
#8	0.483 ± 0.187	**0.492** ± 0.172
#9	0.817 ± 0.186	**0.850** ± 0.200
Mean	0.661 ± 0.146	**0.679** ± 0.152

TABLE III. COMPARISON OF ROC-AUC SCORES FOR THE GLOBAL CONVOLUTIONAL AUTOENCODER (CAE) AND LSTM AUTOENCODER (LSTMAE) UNDER PER-PATIENT AND GLOBAL NORMALIZATION SCHEMES.

Norm.	ROC-AUC	
	CAE	LSTMAE
Per-Patient	0.618 ± 0.023	**0.640** ± 0.031
Global	0.633 ± 0.033	**0.648** ± 0.031

fed into two distinct decoders, one for each modality, trained to reconstruct their respective inputs. During inference, the input reconstruction error is used as the anomaly score, with higher reconstruction errors corresponding to relapsing states.

Audio Branch: Regarding the audio branch, we experimented with both a Convolutional Autoencoder and an LSTM Autoencoder. The Convolutional Autoencoder [12] compresses 128×64 mel-spectrograms through 4 downsampling blocks, each consisting of a ReLU-equipped 2D-Convolution layer and a Max Pooling layer. To reconstruct its input, it applies 4 consecutive convolutional upsampling blocks upon the latent representation, each including an Upsampling layer, a 2D-Convolution layer and a ReLU activation function, followed by an 1-channel 2D-convolution that restores the original dimensions of the mel-spectrogram. The various architectural parameters are the same as in [12].

In the case of the LSTM Autoencoder, the encoder consists of an LSTM layer with 64 units, followed by Layer Normalization, Leaky ReLU activation, and a Dropout layer. The encoded sequence is then flattened, and mapped to a low-dimensional latent space through a Dense layer (of 64 neurons). Conversely, the decoder expands the latent representation with a Dense layer, reshapes it into a $64 \cdot 64$ sequence format, and uses another LSTM layer, symmetric to the encoder, followed by a Time-Distributed layer to reconstruct the original spectrogram.

Physiological Branch: In the physiological branch, the Convolutional Autoencoder, adapted from the best-performing architecture in the e-Prevention study [22], compresses 96×10 physiological feature tensors through 4 downsampling blocks, each consisting of a 1D-Convolutional layer with Batch Normalization, a Leaky ReLU activation, and a Max Pooling layer. The decoder reconstructs the original input using 4 upsampling blocks, each comprising an Upsampling layer followed by a 1D-Convolutional layer with Batch Normalization and a Leaky ReLU activation, and concludes with a Dense layer with a linear activation that restores the original dimensions.

Data Alignment: Since the proposed framework operates on pairs of audio and physiological data, defining a strategy to sample those pairs is of profound importance. Initially, we created pairs by aligning the physiological data to the exact dates of the audio interviews sessions. However, given the requirement for a sufficient number of paired samples across many sessions, this naive approach resulted in an insufficiently sized dataset of only 4 patients. Thus, we relaxed this strict alignment criterion by pairing each interview with physiological data collected within defined time windows ($\pm$ 7 days) around the interview dates, thereby expanding our dataset with additional paired samples; that is, for each interview, a spectrogram is randomly coupled with an 8-hour tensor of physiological data. Under these criteria, data from 7 patients and 158 sessions are incorporated in the dataset.

IV. EXPERIMENTAL SETUP

In our analysis, we employed a common training and evaluation pipeline across both unimodal and multimodal frameworks. Following previous studies [22], [12], we used two experimental setups: **personalized**, where separate models were trained on each patient's data, and **global**, where models were trained on combined data from all patients. In the global experiments, we experimented with i) per-patient normalization by normalizing each patient's data independently, and ii) global normalization by normalizing all patients' data together. Models were trained exclusively on clean data using 5-fold cross-validation and evaluated on clean and anomalous (pre-relapse, relapse) data. Each fold's clean data was split into training, validation, and testing sets (3:1:1 ratio), ensuring data from the same interview date remained within one set to prevent session-wise overfitting. Models were implemented in Keras, with a maximum training duration of 200 epochs and a batch size of 8; early stopping was applied by monitoring the validation loss, with a patience of 10 epochs. We used the Adam optimizer with a learning rate of 3e-4 for the convolutional models and 1e-3 for the LSTM model, and the Mean Squared Error (MSE) as the loss function. For multimodal frameworks, loss weights per branch were experimentally adjusted to balance modality reconstruction quality.

The performance of all models was evaluated on a per-session basis, where anomaly scores from all samples within a session were aggregated into a single score, using reconstruction MSE as the anomaly score. We assessed the models' capability to differentiate between clean and anomalous states using the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). For joint autoencoder models, we initially compared the anomaly scores of the audio and physiological branches individually to their unimodal baselines. Subsequently, we calculated a combined anomaly score, employing a weighted sum of the branches' scores with the same weights used during training, to evaluate overall performance compared to both unimodal baselines. We note that, for personalized setups, we report on the macro-average ROC-AUC score across all patients, whereas for global setups, we use the per-session average ROC-AUC score across sessions.

V. RESULTS & DISCUSSION

Audio Model Comparison: We first compare the performance of the audio autoencoders used in our joint framework, under a unimodal setting. Tables II and III report the results for the personalized and global experiments, respectively, for

TABLE IV. COMPARISON OF ROC-AUC SCORES FOR UNIMODAL MODELS, BRANCHES, AND THE OVERALL COMBINED PERFORMANCE OF THE JOINT AUTOENCODER MODELS ACROSS EXPERIMENTAL SETUPS AND NORMALIZATION SCHEMES.

Exp. Setup	Audio Model	Audio		Bio		Combined
		Unimodal	Branch	Unimodal	Branch	
Personalized	CAE	0.598±0.064	**0.643±0.079**	0.557±0.136	**0.629±0.120**	**0.652±0.170**
	LSTMAE	0.624±0.143	**0.650±0.143**		**0.597±0.151**	**0.627±0.163**
Per-Patient	CAE	0.603±0.059	**0.614±0.048**	0.553±0.029	**0.555±0.032**	0.582±0.028
	LSTMAE	0.632±0.032	0.612±0.042		**0.576±0.033**	0.603±0.043
Global	CAE	0.600±0.060	**0.607±0.053**	0.543±0.050	**0.572±0.039**	0.598±0.046
	LSTMAE	0.617±0.022	**0.629±0.051**		**0.556±0.037**	0.612±0.046

TABLE V. COMPARISON OF ROC-AUC SCORES OF THE UNIMODAL MODELS AND EACH BRANCH OF THE JOINT CONVOLUTIONAL MODEL WITH THE OTHER DISABLED (JOINT UNIMODAL) AND ENABLED (JOINT BRANCH) RESPECTIVELY, FOR BOTH PERSONALIZED AND GLOBAL EXPERIMENTAL SETUPS AND NORMALIZATION SCHEMES.

Modality	Exp. Setup	ROC-AUC		
		Unimodal	Joint Unimodal	Joint Branch
Audio	Personalized	0.598±0.064	0.620±0.108	0.643±0.079
	Per-Patient	0.603±0.059	0.607±0.093	0.614±0.048
	Global	0.600±0.060	0.602±0.056	0.607±0.053
Bio	Personalized	0.557±0.136	0.602±0.108	0.629±0.120
	Per-Patient	0.553±0.029	0.555±0.040	0.555±0.032
	Global	0.543±0.050	0.550±0.057	0.572±0.039

the Convolutional Autoencoder (CAE) and the LSTM Autoencoder (LSTMAE), with superior ROC-AUC scores highlighted in bold. In the personalized experiments, the LSTMAE model outperformed the CAE for nearly all patients, with the exception of Patient #5, raising the average ROC-AUC from 0.661 to 0.679. Notably, in both models, we observe comparatively lower ROC-AUC scores for Patients #2 and #8, presumably due to the low severity of their relapses or limited training data. Similarly, in the global experiments, the LSTMAE achieved superior ROC-AUC scores compared to the CAE under both per-patient and global normalization schemes. These results indicate that the LSTMAE's ability to capture temporal dependencies enhances its ability to distinguish between normal and anomalous states, making it an effective alternative for relapse detection in spontaneous speech data. Finally, we note that due to the differences in dataset selection, direct comparison with other research works is not possible; however, these results are comparable to those reported at the literature, achieved with [12] or without [4] the e-Prevention dataset.

Fusion Scheme Performance: In Table IV, we present the results for the Joint Convolutional Autoencoder and LSTM-Convolutional Autoencoder, comparing the audio and physiological (bio) unimodal baselines to i) their corresponding branches in the joint framework and ii) the combined ROC-AUC score of the joint models, across both personalized and global experimental setups with per-patient and global normalization; bolded values indicate where the branches outperform their unimodal counterparts or the joint model outperforms both. We observe that while the audio modality is generally stronger than the physiological modality, the branches consistently outperform their unimodal counterparts in both experimental setups, demonstrating that the complementary information from the added modality enhances relapse prediction. In the personalized experiments, the overall performance of the joint models exceeds both unimodal baselines, whereas in the global experiments the combined performance is enhanced primarily over the physiological baseline. Thus, we conclude that the multimodal fusion approach is more effective in a personalized setting for detecting relapse from individual patients' speech and physiological signals.

An advantage of the feature-level fusion scheme is that, in contrast to late fusion, it can also operate when one of the two modalities is missing. Thus, we disable each branch of the Joint Convolutional Autoencoder by zeroing its input, and proceed to the evaluation as before. In Table V, we present the results obtained from the branch disabling experiments, reporting on i) the unimodal models, ii) the joint models with one branch disabled, and iii) the joint models with both modalities available. Across both modalities, a progressive improvement is evident; while the jointly trained model performs better under the availability of both audio and physiological modalities, the results obtained after disabling one branch of the joint model are superior to those of the respective unimodal model. Notably, in the global experiments, the progressive improvement is more moderate than in the personalized setting, which is consistent with the results discussed earlier. Overall, these findings confirm that the joint framework effectively leverages complementary information from both speech and physiological signals, enhancing relapse detection, and highlighting the value of the feature-level fusion approach.

VI. CONCLUSION

In this work, we proposed and evaluated advanced autoencoder architectures for improved relapse detection in patients with psychotic disorders using spontaneous speech and physiological signals. We introduced LSTM-based autoencoders and compared them to convolutional ones, demonstrating that the LSTM architecture's ability to capture temporal dependencies in speech data makes it an effective and robust alternative for relapse detection. We also proposed a multimodal joint autoencoder framework employing a feature-level fusion approach, which proved effective and significantly enhanced accuracy in relapse prediction compared to unimodal baselines. These frameworks were particularly effective in personalized setups, highlighting the importance of individual patient monitoring. Furthermore, ablation studies underscored the complementary nature of audio and physiological modalities. For future work, we are interested in exploring pretraining and fine-tuning strategies tailored specifically to individual patients, as well as integrating additional modalities such as video or text. Another interesting avenue for exploration is user grouping or clustered federated learning methods to further enhance the model's predictive performance, whereas transfer learning from larger audio/physiological signal datasets, as well as data augmentations [30], could be applied to overcome dataset size limitations.

Acknowledgments: This project is funded by the European Union under Horizon Europe (grant No. 101136568 - project HERON)

Funded by the European Union

REFERENCES

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

[2] M. Bauer, T. Wilson, K. Neuhaus, J. Sasse, A. Pfennig, U. Lewitzka, P. Grof, T. Glenn, N. Rasgon, T. Bschor, and P. C. Whybrow, "Self-reporting software for bipolar disorder: Validation of chronorecord by patients with mania," *Psychiatry Research*, vol. 159, no. 3, pp. 359–366, 2008.

[3] R. Wang, M. S. H. Aung, S. Abdullah, R. Brian, A. T. Campbell, T. Choudhury, M. Hauser, J. Kane, M. Merrill, E. A. Scherer, V. W. S. Tseng, and D. Ben-Zeev, "Crosscheck: toward passive sensing and detection of mental health changes in people with schizophrenia," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery, 2016, p. 886–897.

[4] J. Gideon, K. Matton, S. Anderau, M. G. McInnis, and E. Mower Provost, "When to intervene: Detecting abnormal mood using everyday smartphone conversations," *IEEE Transactions on Affective Computing (Preprint)*, 2019.

[5] I. Barnett, J. Torous, P. Staples, L. Sandoval, M. Keshavan, and J. P. Onnela, "Relapse prediction in schizophrenia through digital phenotyping: A pilot study," *Neuropsychopharmacology*, 2018.

[6] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. of the 22nd ACM SIGKDD Int'ernationa'l Conf. on Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, CA, USA, 2016.

[7] A. Maxhuni, A. Muñoz-Meléndez, V. Osmani, H. Perez, O. Mayora, and E. F. Morales, "Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients," *Pervasive and Mobile Computing*, vol. 31, pp. 50–65, Feb. 2016.

[8] A. Ikäheimonen, N. Luong, I. Baryshnikov, R. Darst, R. Heikkilä, J. Holmen, A. Martikkala, K. Riihimäki, O. Saleva, E. Isometsä, and T. Aledavood, "Predicting and monitoring symptoms in diagnosed depression using mobile phone data: An observational study," *medRxiv Preprint*, 2024.

[9] D. A. Adler, D. Ben-Zeev, V. W.-S. Tseng, J. M. Kane, R. Brian, A. T. Campbell, M. Hauser, E. A. Scherer, and T. Choudhury, "Predicting early warning signs of psychotic relapse from passive sensing data: An approach using encoder-decoder neural networks," *JMIR mHealth and uHealth*, vol. 8, no. 8, p. e19962, 2020.

[10] M. Panagiotou, A. Zlatintsi, P. P. Filntisis, A. Roumeliotis, N. Efthymiou, and P. Maragos, "A comparative study of autoencoder architectures for mental health analysis using wearable sensors data," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 1258–1262.

[11] 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. 695–700, 2018.

[12] C. Garoufis, A. Zlatintsi, P. P. Filntisis, N. Efthymiou, E. Kalisperakis, V. Garyfalli, T. Karantinos, L. Mantonakis, N. Smyrnis, and P. Maragos, "An unsupervised learning approach for detecting relapses from spontaneous speech in patients with psychosis," in *IEEE EMBS Int'l Conf. on Biomedical and Health Informatics (BHI)*, 2021.

[13] 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*, 2018.

[14] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of Biomedical Informatics*, vol. 83, pp. 103–111, 2018.

[15] S. Walther and V. A. Mittal, "Motor system pathology in psychosis," *Current Psychiatry Reports*, vol. 19, p. 97, Oct. 2017.

[16] M. Faurholt-Jepsen, J. Busk, M. Frost, M. Vinberg, E. M. Christensen, O. Winther, J. E. Bardram, and L. V. Kessing, "Voice analysis as an objective state marker in bipolar disorder," *Translational Psychiatry*, vol. 6, p. e856, Jul. 2016.

[17] A. Othmani and A. O. Zeghina, "A multimodal computer-aided diagnostic system for depression relapse prediction using audiovisual cues: A proof of concept," *Healthcare Analytics*, vol. 2, p. 100090, 2022.

[18] R. Flores, M. Tlachac, E. Toto, and E. Rundensteiner, "Audiface: Multimodal deep learning for depression screening," *Proc. of Machine Learning Research*, vol. 182, pp. 1–22, 2022.

[19] G. Lam, D. Huang, and W. Lin, "Context-aware deep learning for multi-modal depression detection," in *Proc. of the IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[20] F. F. de Almeida, K. R. T. Aires, A. C. B. Soares, L. de Sousa Britto Neto, and R. de Melo Souza Veras, "Multimodal fusion for depression detection assisted by stacking deep neural networks," *IEEE Journal*, 2024.

[21] Y. Wang, Z. Wang, C. Li, Y. Zhang, and H. Wang, "Online social network individual depression detection using a multitask heterogeneous modality fusion approach," *Information Sciences*, vol. 609, pp. 727–749, 2022.

[22] A. Zlatintsi, P. P. Filntisis, C. Garoufis, N. Efthymiou, P. Maragos, A. Menychtas, I. Maglogiannis, P. Tsanakas, T. Sounapoglou, E. Kalisperakis *et al.*, "E–prevention: Advanced support system for monitoring and relapse prevention in patients with psychotic disorders analyzing long–term multimodal data from wearables and video captures," *Sensors*, vol. 22, no. 19, p. 7544, Oct. 2022.

[23] P. Mobtahej, X. Zhang, M. Hamidi, and J. Zhang, "An lstm-autoencoder architecture for anomaly detection applied on compressors audio data," *Computational and Mathematical Methods*, 2022.

[24] G. Papathanasiou, G. Georgoudis, M. Papandreou, P. Spyropoulos, D. Georgakopoulos, V. Kalfakakou, and A. Evangelou, "Reliability measures of the short international physical activity questionnaire (ipaq) in greek young adults," *Hellenic Journal of Cardiology*, vol. 50, pp. 283–294, 2009.

[25] 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. Advances in Information and Communication Technology (AIAI)*, 2020.

[26] 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, Jun. 1987.

[27] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. of the Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, AB, Canada, Apr. 2018.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Hawaii, USA, Dec. 2011.

[29] J. D. Scargle, "Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data," *The Astrophysical Journal*, vol. 263, pp. 835–853, 1982.

[30] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech 2019*, Graz, Austria, 2019.