Uncertainty-Driven Anomaly Detection for Psychotic Relapse Using Smartwatches: Forecasting and Multi-Task Learning Fusion

N. Tsalkitzis[2], P. P. Filntisis[1,3], P. Maragos[1,2,3] and N. Efthymiou[1,3]
[1]Robotics Institute, Athena Research and Innovation Center, Greece
[2]School of ECE, National Technical University of Athens, Greece
[3]HERON - Hellenic Robotics Center of Excellence, Athens, Greece
{nefthymiou, pflntisis}@athenarc.gr, {maragos}@cs.ntua.gr, nikostsalkitzhs@gmail.com

***Abstract*—Digital phenotyping enables continuous passive monitoring of behavior and physiology, offering a promising paradigm for early detection of psychotic relapse. In this work, we develop and systematically study two smartwatch-based frameworks for daily relapse detection. The first forecasts cardiac dynamics and flags deviations between predicted and observed features as indicators of abnormality. The second adopts a multi-task formulation that fuses sleep with motion and cardiac-derived signals, learning time-aware embeddings and predicting measurement timing. Both pipelines use Transformer encoders and output a daily anomaly score, derived from predictive uncertainty estimated via an ensemble of multilayer perceptrons to improve robustness to real-world wearable variability. While each framework independently demonstrates strong predictive power, we show that they capture complementary physiological signatures. Consequently, we propose a late-fusion strategy that synergistically combines the anomaly signals from both architectures into a unified decision score. We benchmark our methodology on the 2nd e-Prevention Grand Challenge dataset, where our fused model achieves an 8% relative improvement over the competition-winning baseline. Our results, supported by extensive ablation studies, suggest that the integration of diverse digital phenotypes, cardiac, motion, and sleep, is essential for the high-fidelity detection of psychotic relapse in real-world settings.**

***Index Terms*—relapse prediction, digital phenotyping, transformer encoders, uncertainty estimation, smartwatch**

I. Introduction

Digital phenotyping has increasingly emerged as a valuable approach in medical applications, with particularly strong momentum in psychiatry [1], [2]. By leveraging unobtrusively digital data during everyday life [3], [4], clinicians can enrich standard assessments with additional context about an individual's mental state. Because these measurements are collected continuously in naturalistic settings, they can reveal both overt behavioral changes and subtle shifts that may precede clinical deterioration [5]. This capability supports earlier detection of clinically meaningful changes, including signals of impending psychotic relapse, and allows for personalized interventions that augment traditional care [6].

Relapse detection is typically framed as supervised learning when labels are abundant, or unsupervised when events are rare. While [7] uses a bidirectional LSTM for weekly probability estimates, the scarcity of real-world relapse data often favors unsupervised models that learn a "remission baseline." For instance, [8] treats identity misclassification as a proxy for clinical change, while [9] uses a Convolutional Autoencoder to extract latent features from 4-hour 2D profiles, deriving relapse scores from sleep-specific reconstruction errors. More complex architectures include multi-branch CNNs with multi-head attention to visualize 24-hour activity patterns [10], and Transformer-based encoder-decoders that leverage smartphone features to flag individualized behavioral anomalies [4].

Recently, during the 2nd e-Prevention Grand Challenge on psychotic relapse detection a biosignal dataset [11] was introduced. SoTA work [12] in the challenge employs a patient-specific Transformer-based unsupervised anomaly detection framework that learns an individual's typical daily routine by predicting the timestamp of wearable biosignal measurements using non-relapse data, deriving daily anomaly scores via an ensemble of MLP heads to stabilize prediction error. Other approaches, such as [13], utilize separate autoencoders for multimodal data combined with Elliptic Envelopes to calculate relapse probabilities, while [14] leverages a Transformer encoder with cross-entropy and prototype losses to identify outlier days.

Building on these foundations, we present a novel Transformer-based framework for modeling wearable-derived physiological signals taking advantage of both cardiac variability and sleep-wake regularity that are associated with within-subject fluctuations in psychotic symptom severity, and departures from an individual's typical patterns may therefore provide early evidence of relapse risk [15]. Specifically, we study two approaches, one for forecasting future cardiac features and the second learning time-aware and sleep embeddings via a multi-task Transformer encoder, where an MLP ensemble estimates predictive uncertainty to derive the final anomaly score. Their fusion results in a robust framework for enabling early detection of relapses.

In this paper, our contributions are summarized as follows:

- We develop two smartwatch-based Transformer architectures for daily psychotic relapse detection: one for cardiac feature forecasting, and another that leverages multi-task

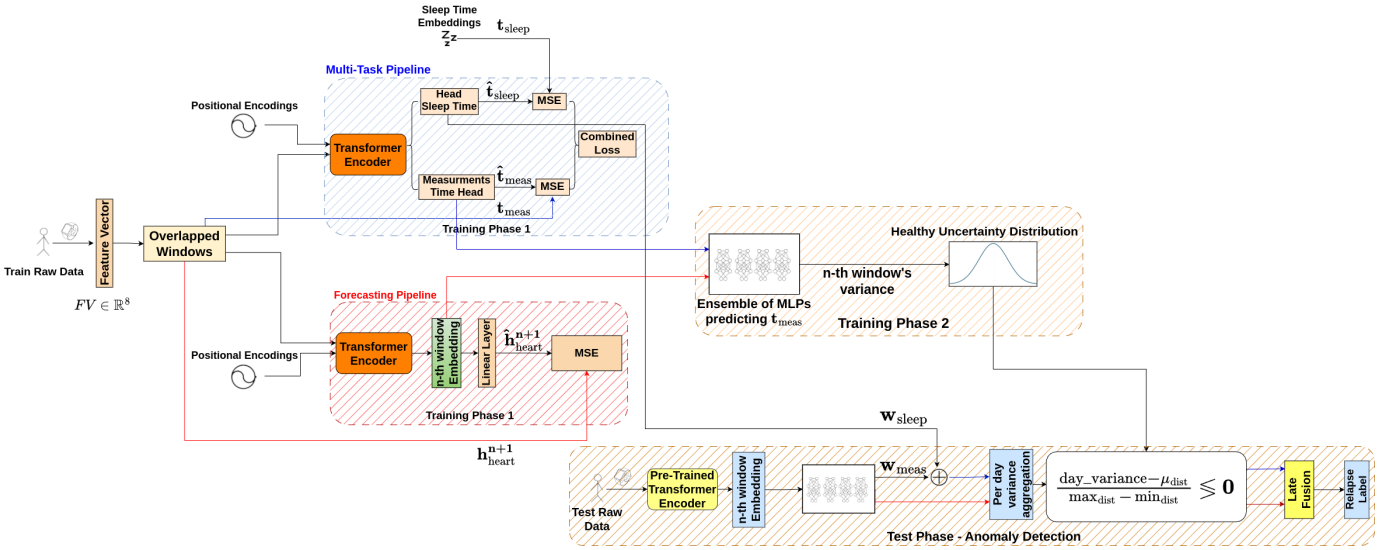

Fig. 1: The proposed transformer-based relapse and anomaly detection framework where windowed wearable features are encoded and then routed through one of two alternative learning paths either a **forecasting** pipeline shown in red or a **multi-task** sleep and measurement time pipeline shown in blue. The resulting predictions are used to derive uncertainty estimates via an MLP ensemble which are aggregated and late fused during testing to produce the final relapse label.

learning with sleep-based, time-aware embeddings. Both models deliver strong performance on the e-Prevention benchmark, achieving a 7% improvement over the SoTA work.

- We systematically investigate late-fusion strategies for combining the two models and show that jointly leveraging their anomaly scores yields improved performance, reaching an 8% gain.

II. Dataset

For our study, we use the dataset from Track 2 of the ICASSP 2024 e-Prevention Grand Challenge [11], part of the e-Prevention dataset [16], which consists of smartwatch-derived biosignals from eight patients with psychotic disorders, such as schizophrenia and bipolar disorder, all of whom experienced at least one psychotic relapse during the recording period. The raw dataset comprises continuous recordings from multiple modalities, including accelerometer and gyroscope signals, RR intervals, heart rate, step counts, and sleep-related activity. For analysis, each participant's recordings are segmented into daily windows, and the raw measurements are aggregated into 5-minute intervals to reduce the impact of sensor-level noise on downstream classification, consistent with the considerations reported in [17]. Subsequently, we carefully extract the following features: the acceleration norm and gyroscope norm, the mean heart rate, and heart-rate-variability descriptors computed from the RR-interval series, including the mean RR interval, the Root Mean Square of Successive Differences (RMSSD), the Standard Deviation of Normal-to-Normal intervals (SDNN), and the high-frequency Lomb–Scargle power. In addition, we incorporate time embeddings for the measurement timestamps, as well as embeddings for the sleep onset and wake-up times.

On average, the training split contains approximately 200 days of recordings, while the validation and test splits comprise about 87 and 85 days [11], respectively. Importantly, the training data include remission periods only, whereas both the validation and test data contain samples from remission as well as relapse periods. After extracting features for each participant, we segment each day into overlapping windows formed over the 288 five-minute intervals. The number of windows per day is given by

$$N_{\text{win}} = \left\lfloor \frac{288 - \textit{window\_size}}{\textit{stride}} \right\rfloor + 1, \tag{1}$$

where 288 denotes the number of five-minute intervals in a day, *stride* controls the displacement of the sliding window, and *window_size* denotes the window length.

III. Framework Architecture

The proposed architectures, both based on Transformer encoders, are presented in Fig. 1. In the forecasting pipeline, the encoder is used to predict future trajectories of cardiac features, whereas the multi-task model jointly learns time-aware and sleep-related embeddings. Subsequently, an ensemble of multilayer perceptrons is employed to quantify predictive uncertainty, which is then aggregated into an uncertainty score and, ultimately, utilized to derive the final anomaly score. This approach leverages the fact that shifts in cardiac and sleep–wake patterns often track with psychotic symptom severity and provide early warnings of impending relapse.

A. Forecasting Architecture

For **the forecasting-based pipeline**, the input at each five-minute timestep consists of the acceleration and gyroscope norm, the five cardiac features and the count of steps. These sequences are passed through a Transformer encoder, which produces a latent representation per timestep. Average pooling over the window yields a single embedding per window, which is subsequently used to predict the cardiac-feature vector at the next timestep. Formally, let $\mathbf{x}_t \in \mathbb{R}^d$ denote the input feature

vector at timestep t, and let $\mathbf{c}_t \in \mathbb{R}^5$ denote the corresponding vector of cardiac features. For a window of length L starting at index s, define the input segment $\mathbf{X}_s = [\mathbf{x}_s, \ldots, \mathbf{x}_{s+L-1}]$. Given $\mathbf{X}_s$, the encoder produces a sequence of hidden states, $\mathbf{H}_s = \mathrm{Enc}(\mathbf{X}_s)$, where $\mathbf{H}_s = [\mathbf{H}_{s,0}, \ldots, \mathbf{H}_{s,L-1}]$ and $\mathbf{H}_{s,i}$ denotes the encoder output at offset i within the window. We obtain a fixed-dimensional representation for the entire window by mean pooling,

$$\mathbf{z}_s = \frac{1}{L} \sum_{i=0}^{L-1} \mathbf{H}_{s,i}, \tag{2}$$

and pass this window-level embedding to a predictor head $\mathbf{g}$ to forecast the next-step cardiac features, $\hat{\mathbf{c}}_{s+L} = g(\mathbf{z}_s)$. The model is trained using a one-step-ahead forecasting objective over a batch $\mathcal{B}$ of windows,

$$\mathcal{L}_{\text{forecast}} = \frac{1}{\mathcal{B}} \sum_{s \in \mathcal{B}} \|\hat{\mathbf{c}}_{s+L} - \mathbf{c}_{s+L}\|_2^2 . \tag{3}$$

In a second training phase, we reuse the same encoder embedding $\mathbf{z}_s$, but replace the single predictor $\mathbf{g}$ with an ensemble of K multilayer perceptrons (MLPs). The learning objective remains the same one-step-ahead cardiac forecasting loss, while diversity is promoted via an online resampling scheme. For each mini-batch, embedding-target pairs are replicated across all K heads, and then, independently per head, a subset of samples is replaced with embedding-target pairs drawn from randomly selected training instances. Let $\{f_k(\cdot)\}_{k=1}^K$ denote the ensemble and let $\tilde{\mathbf{z}}_s^{(k)}$ denote the resampled input to head k. Each head produces a next-step prediction,

$$\hat{\mathbf{c}}_{s+L}^{(k)} = f_k\left(\tilde{\mathbf{z}}_s^{(k)}\right), \qquad k = 1, \ldots, K, \tag{4}$$

and we aggregate predictions using the ensemble mean,

$$\bar{\mathbf{c}}_{s+L} = \frac{1}{K} \sum_{k=1}^{K} \hat{\mathbf{c}}_{s+L}^{(k)}. \tag{5}$$

To quantify predictive uncertainty for window s, we compute the element-wise ensemble variance,

$$\boldsymbol{\sigma}_s^2 = \frac{1}{K} \sum_{k=1}^{K} \left(\hat{\mathbf{c}}_{s+L}^{(k)} - \bar{\mathbf{c}}_{s+L}\right)^{\odot 2}, \tag{6}$$

where $(\cdot)^{\odot 2}$ denotes element-wise squaring and $\boldsymbol{\sigma}_{s,j}^2$ corresponds to the variance of the j-th cardiac feature. We summarize this vector by averaging across the five cardiac-feature dimensions,

$$u_s = \frac{1}{5} \sum_{j=1}^{5} \boldsymbol{\sigma}_{s,j}^2. \tag{7}$$

Finally, the daily variance score is defined as the mean uncertainty across all windows whose start indices fall within day d,

$$U_d = \frac{1}{\mathcal{W}_d} \sum_{s \in \mathcal{W}_d} u_s, \tag{8}$$

where $\mathcal{W}_d$ representing the number of windows in day d.

Using the remission-only training split, we construct a patient-specific (healthy) uncertainty distribution from the day-level scores $\{U_d\}$ and summarize it via its empirical minimum, maximum, and mean values. During inference, the proposed architecture produces a day-level uncertainty score U_d for each day d. For remission days, this score is expected to remain close to the mean of the healthy distribution, whereas relapse days are characterized by systematic deviations. Accordingly, we define the final anomaly score for day d as the normalized deviation from the healthy mean:

$$A_d = \frac{U_d - \mu_{\text{dist}}}{\max_{\text{dist}} - \min_{\text{dist}}}, \tag{9}$$

where μ_{dist}, $\max_{\text{dist}}$, and $\min_{\text{dist}}$ denote the mean, maximum, and minimum of the patient-specific healthy uncertainty distribution, respectively. Finally, we apply a fixed decision rule by thresholding the anomaly score:

$$\hat{y}_d = \begin{cases} 1, & \text{if } A_d > \tau \quad \text{(relapse)}, \\ 0, & \text{otherwise} \quad \text{(remission)}. \end{cases} \tag{10}$$

B. Multi-Task Architecture

In contrast, for **the multi-task learning approach** we build upon and extend the winning architecture of the challenge [12]. The model receives two input streams. The first comprises physiological and activity-related signals, including cardiac features, the norms of the inertial sensors, and step counts. The second stream consists of two pairs of time embeddings that encode sleep onset and wake-up time. Training is performed with a multi-task objective that combines (i) the prediction of measurement-time embeddings from the observed signals and (ii) an auxiliary sleep-prediction head that estimates sleep-related time embeddings. For uncertainty estimation and the construction of the patient-specific healthy distribution, we exclude the sleep branch and rely solely on uncertainty derived from the measurement-time embedding task, since sleep-related measurements are comparatively noisy in the available data [12]. During inference, we compute a combined day-level variance score by fusing the variance from the sleep head with the uncertainty associated with the measurement-time embeddings, using weights of 0.3 and 0.7, respectively. The resulting score is then normalized using the summary statistics of the healthy distribution to obtain the final anomaly score.

C. Late Fusion

To leverage the complementary strengths of the heart-forecasting and the sleep/circadian architectures, we investigated *late fusion* at the score level. Each model produces a continuous day-level anomaly score and instead of thresholding the two systems independently, we fuse these continuous scores into a single risk score and apply thresholding only at the end. This retains more information for ranking-based evaluation and allows the two modalities to compensate for one another. Our primary fusion mechanism is a convex combination of the two anomaly scores:

$$\text{fused\_score} = \alpha \cdot A_{D,\text{heart}} + (1 - \alpha) \cdot A_{D,\text{sleep}} \tag{11}$$

where $\alpha \in [0, 1]$ controls the relative contribution of each modality. The value of α was selected on the validation set, via grid search, to maximize detection performance. In addition to weighted averaging, we explored max and min fusion:

$$\text{fused\_score} = \{\max, \min\}(A_{D,\text{heart}},\ A_{D,\text{sleep}}) \tag{12}$$

Max fusion acts as an *OR*-style rule, producing a high fused score when *either* model strongly signals abnormality. This is beneficial when relapses may manifest predominantly in one modality (e.g., sleep disruption without marked cardiac changes, or vice versa), thereby improving sensitivity. Conversely, min fusion behaves like an *AND*-style rule, yielding a high fused score only when *both* modalities are simultaneously anomalous. This can reduce false positives by requiring cross-modal agreement, at the cost of lower recall when a relapse signal is strong in only one of the two models.

IV. Experimental Analysis

All models were trained for 50 epochs with a learning rate of 10^{-3}, weight decay 5×10^{-4}, and a batch size of 16. We used an ensemble of five MLP predictors. Performance was evaluated using AUROC (Area Under the ROC Curve), AUPRC (Area Under the Precision-Recall Curve), and their arithmetic mean (AVG), reported as the average over 10 independent runs.

Positional Embeddings: Since all methods are Transformer-based we compared positional encodings per task. Table I showcases that ALiBi [18] is most effective for sleep multi-task modeling because its decaying attention bias, where scores are penalized by token distance, inherently suits sleep architecture. This method captures local context such as stage transitions while staying invariant to absolute shifts between nights, matching the relative and variable nature of sleep recordings. For cardiac forecasting sinusoidal embeddings perform best by providing an absolute temporal anchor that preserves consistent periodic signals like circadian and time-of-day effects. RoPE [19] is competitive but its rotation-based relative encoding yields lower performance than the best method for each task.

TABLE I: Positional Embedding Performance. Best per model/metric in **bold**; second-best is underlined.

Model	Positional encoding	AUROC	AUPRC	AVG
Forecasting	RoPE [19]	0.562 ± 0.013	0.549 ± 0.011	0.556 ± 0.011
	ALiBi [18]	0.547 ± 0.031	0.531 ± 0.033	0.539 ± 0.032
	sinusoidal	$\mathbf{0.565 \pm 0.024}$	$\mathbf{0.552 \pm 0.022}$	$\mathbf{0.559 \pm 0.023}$
Multi-task	RoPE [19]	0.504 ± 0.012	0.606 ± 0.014	0.555 ± 0.010
	ALiBi [18]	$\mathbf{0.505 \pm 0.009}$	$\mathbf{0.613 \pm 0.012}$	$\mathbf{0.559 \pm 0.010}$
	sinusoidal	0.502 ± 0.013	0.602 ± 0.022	0.552 ± 0.014

Window Hyperparameters Ablation: The number of windows is governed by two key parameters, namely the stride and the window size. To determine an appropriate stride value for both models, we conducted a sensitivity analysis, the results of which are reported in Table II.

TABLE II: Stride sensitivity (mean $\pm$ std). Best per model/metric in **bold**; second-best is underlined.

Model	Stride	AUROC	AUPRC	AVG
Forecasting	12 (1 hr)	**0.568 ± 0.022**	**0.554 ± 0.019**	**0.561 ± 0.019**
	24 (2 hr)	0.538 ± 0.020	0.551 ± 0.015	0.545 ± 0.017
	36 (3 hr)	0.542 ± 0.019	0.547 ± 0.019	0.545 ± 0.018
	48 (4 hr)	0.535 ± 0.029	0.522 ± 0.027	0.528 ± 0.028
Multi-task	12 (1 hr)	0.506 ± 0.009	0.615 ± 0.011	0.560 ± 0.009
	24 (2 hr)	**0.506 ± 0.005**	**0.617 ± 0.008**	**0.562 ± 0.006**
	36 (3 hr)	0.503 ± 0.013	0.594 ± 0.024	0.548 ± 0.017
	48 (4 hr)	0.496 ± 0.010	0.574 ± 0.025	0.535 ± 0.016

With the stride results above, a one-hour stride yields the best performance for both architectures. Keeping this stride fixed, a subsequent window-size sensitivity analysis indicates that a two-hour window (24 timesteps) is optimal, corresponding to a 50% overlap between consecutive windows. This choice provides a favorable trade-off between data efficiency and temporal context: it increases the number of training samples while capturing meaningful short-term dynamics, without overly long windows that may dilute relapse-related patterns and introduce additional noise or nonstationarities.

Threshold τ Ablation: We conducted an ablation study on the validation set to examine how the relapse decision threshold τ affects performance. This threshold acts as a decision cutoff. It represents the minimum daily risk score required for a day to be labeled as a relapse. Decreasing τ makes relapse detection more permissive, whereas increasing τ makes it more conservative. Across both modalities, we observed that negative thresholds generally provide a better balance between ranking ability and precision under class imbalance. In addition, the two models show distinct strengths. The forecasting model tends to achieve stronger discrimination (higher AUROC), while the multi-task model typically yields better early-retrieval behavior (higher AUPRC), with this contrast being especially evident around $\tau = 0$. Therefore, we select $\tau = -0.1$ and $\tau = -0.2$, respectively for the two architectures.

V. Results & Discussion

To assess which model yields the strongest performance, we further benchmark our methods against both the baseline and first two winning architectures of the challenge, with the comparative results summarized in Table III and Figure 2.

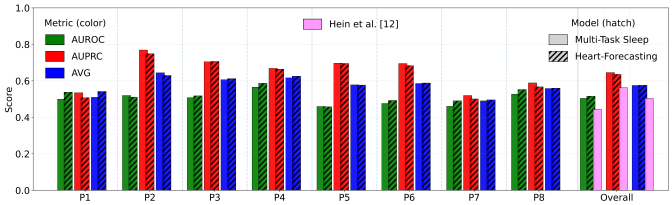

Fig. 2: Patient-wise AUROC, AUPRC, and AVG scores of the proposed forecasting and multi-task models compared against Hein *et al.* [12]

TABLE III: Test-set performance for Track 2 reference/proposed models and score-fusion variants (best in **bold**; second best underlined).

Group	Method	AUROC	AUPRC	AVG
Single	Baseline Approach	0.495	0.379	0.437
	Hein [12]	0.444	0.563	0.504
	Mallo-Ragolta [13]	0.493	0.505	0.499
	Multi-task Model	0.505	**0.646**	0.575
	Forecasting-based Model	**0.516**	0.636	**0.576**
Fusion	Weighted (α=0.7, τ=−0.1)	0.504	0.664	**0.584**
	Max (τ=−0.1)	0.501	**0.667**	**0.584**
	Min (τ=−0.1)	**0.510**	0.653	0.581

Based on the results above, both proposed models outperform the 2nd challenge's winning architecture of Hein *et*

al. [12] by approximately 7% in AVG, supporting the benefit of uncertainty-driven anomaly scoring.

Per-person analysis further confirms that the two architectures are functionally complementary. The sleep-based model more often achieves higher AUPRC (e.g., P2 and P4–P8), suggesting improved identification of rare relapse days, while the heart-based model more frequently yields higher AUROC (e.g., P3, P4, and P6–P8), indicating stronger overall separability between relapse and non-relapse days. This distinction is clinically relevant under pronounced class imbalance: AUPRC is typically more informative because it emphasizes performance on the relapse class. The heterogeneity in per-patient AVG (sleep higher for P2 and P5; heart higher for P3, P4, and P6–P8) further suggests that relapse signatures are patient-specific and modality-dependent, motivating individualized balancing of modalities.

Fusion improves robustness by leveraging strengths from both pipelines. Across fusion rules (with α and τ selected on the validation set), max fusion attains the highest AUPRC and ties for the best AVG, while weighted fusion matches the top AVG with slightly lower AUPRC. In contrast, min fusion achieves the highest AUROC but the lowest AUPRC, consistent with better global separability yet fewer high-confidence relapse predictions and reduced recall. Overall, late fusion provides the strongest final scheme, and patient-specific fusion weights may yield further gains.

VI. Conclusion

We develop two personalized, unsupervised relapse detectors from smartwatch digital phenotypes that compute day-level anomaly scores via ensemble-based predictive uncertainty: a cardiac forecasting model and a sleep/time multi-task model. Both models surpass the challenge-winning baseline, capturing distinct aspects of relapse risk (cardiac dynamics primarily improve separability, while sleep/time structure is especially informative under severe class imbalance). Late score-level fusion integrates these signals into our final scheme and achieves the strongest overall results (AVG = 0.584; AUPRC up to 0.667), corresponding to an $\sim$8% relative improvement over the winner. Future work will explore early fusion approaches that jointly model cardiac and time/sleep patterns within a unified network to learn cross-modal interactions that may further improve relapse detection, alongside explainability mechanisms to identify which physiological signals drive anomaly scores and improve clinical interpretability.

References

[1] J.-P. Onnela, "Opportunities and challenges in the collection and analysis of digital phenotyping data," *Neuropsychopharmacology*, vol. 46, pp. 45–54, 2021.

[2] T. R. Insel, "Digital phenotyping: a global tool for psychiatry," *World Psychiatry*, vol. 17, no. 3, p. 276, 2018.

[3] J. Huang, Y. Zhao, W. Qu, Z. Tian, Y. Tan, Z. Wang, and S. Tan, "Automatic recognition of schizophrenia from facial videos using 3D convolutional neural network," *Asian Journal of Psychiatry*, vol. 77, p. 103263, 2022.

[4] D. A. Adler, D. Ben-Zeev, V. W. S. Tseng, J. M. Kane, *et al.*, "Predicting early warning signs of psychotic relapse from passive sensing data: an approach using encoder-decoder neural networks," *JMIR mHealth and uHealth*, vol. 8, p. e19962, 2020.

[5] E. Rodriguez-Villa, U. M. Mehta, J. Naslund, D. Tugnawat, *et al.*, "Smartphone Health Assessment for Relapse Prevention (SHARP): a digital solution toward global mental health," *BJPsych Open*, vol. 7, p. e29, 2021.

[6] I. Barnett, J. Torous, P. Staples, L. Sandoval, *et al.*, "Relapse prediction in schizophrenia through digital phenotyping: a pilot study," *Neuropsychopharmacology*, vol. 43, pp. 1660–1666, 2018.

[7] B. Lamichhane, J. Zhou, and A. Sano, "Psychotic relapse prediction in schizophrenia patients using a personalized mobile sensing-based supervised deep learning model," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, pp. 3246–3257, 2023.

[8] N. Efthymiou, G. Retsinas, P. P. Filntisis, C. Garoufis, *et al.*, "From digital phenotype identification to detection of psychotic relapses," in *Proc. ICHI*, 2023.

[9] A. Y. Yan, T. J. Speed, and C. O. Taylor, "Relapse prediction using wearable data through convolutional autoencoders and clustering for patients with psychotic disorders," *Scientific Reports*, vol. 15, p. 18806, 2025.

[10] M. M. Misgar and M. P. S. Bhatia, "Unveiling psychotic disorder patterns: A deep learning model analysing motor activity time-series data with explainable AI," *Biomedical Signal Processing and Control*, vol. 91, p. 106000, 2024.

[11] P. P. Filntisis, N. Efthymiou, G. Retsinas, A. Zlatintsi, *et al.*, "The 2nd E-Prevention Challenge: Psychotic and Non-Psychotic Relapse Detection Using Wearable-Based Digital Phenotyping," in *Proc. ICASSP*, 2024.

[12] A. Hein, S. Gronauer, and K. Diepold, "Patient-specific modeling of daily activity patterns for unsupervised detection of psychotic and non-psychotic relapses," in *ICASSPW*, 2024.

[13] A. Mallol-Ragolta, A. Spiesberger, A. Triantafyllopoulos, and B. Schuller, "Personalised anomaly detectors and prototypical representations for relapse detection from wearable-based digital phenotyping," in *ICASSPW*, 2024.

[14] J. Wu, and M. Tu, "Unsupervised relapse detection using wearable-based digital phenotyping for the 2nd e-prevention challenge, " in *ICASSPW*, 2024.

[15] E. Kalisperakis, T. Karantinos, M. Lazaridi, V. Garyfalli, *et al.*, "Smartwatch digital phenotypes predict positive and negative symptom variation in a longitudinal monitoring study of patients with psychotic disorders," *Frontiers in Psychiatry*, vol. 14, 2023.

[16] A. Zlatintsi, P. P. Filntisis, C. Garoufis, N. Efthymiou, *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, p. 7544, 2022.

[17] G. Retsinas, P. P. Filntisis, N. Efthymiou, and E. Theodosis, *et al.* "Person identification using deep convolutional neural networks on short-term signals from wearable sensors," in Proc *ICASSP*, 2020.

[18] O. Press, N. A. Smith, and M. Lewis, "Train short, test long: Attention with linear biases enables input length extrapolation," in *Proc. ICLR*, 2022.

[19] J. Su, M. Ahmed, Y. Lu, and S. Pan, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

Acknowledgment

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

Funded by
the European Union