From Digital Phenotype Identification To Detection Of Psychotic Relapses

Niki Efthymiou
School of ECE
NTUA
Athens, Greece
nefthymiou@central.ntua.gr

George Retsinas
School of ECE
NTUA
Athens, Greece
gretsinas@central.ntua.gr

Panagiotis P. Filntisis
School of ECE
NTUA
Athens, Greece
filby@central.ntua.gr

Christos Garoufis
School of ECE
NTUA
Athens, Greece
cgaroufis@mail.ntua.gr

Athanasia Zlatintsi
School of ECE
NTUA
Athens, Greece
nzlat@cs.ntua.gr

Emmanouil Kalisperakis
Lab. of Cognitive Neuroscience
Univ. Mental Health RI
Medical School, NKUA
Athens, Greece
mcasper23@hotmail.com

Vasiliki Garyfalli
Lab. of Cognitive Neuroscience
Univ. Mental Health RI
Medical School, NKUA
Athens, Greece
vasiaog@gmail.com

Thomas Karantinos
Lab. of Cognitive Neuroscience
Univ. Mental Health RI
Athens, Greece
tkarantinos@gmail.com

Marina Lazaridi
Lab. of Cognitive Neuroscience
Univ. Mental Health RI
Medical School, NKUA
Athens, Greece
ma.lazaridi@gmail.com

Nikolaos Smyrnis
Lab. of Cognitive Neuroscience
Univ. Mental Health RI
Medical School, NKUA
Athens, Greece
smyrnis@med.uoa.gr

Petros Maragos
School of ECE
NTUA
Athens, Greece
maragos@cs.ntua.gr

***Abstract*—Timely detection of relapses constitutes an important step towards improving the quality of life in patients with psychotic disorders. In this paper, we design a novel framework for discovering indications of psychotic relapses by modeling the digital phenotype of the patients who wear smartwatches. We start by designing deep neural network architectures that can use biosignals for person identification with high discriminatory performance. Then, we show how these networks can be employed to identify indications of psychotic relapses by looking at the per-person misclassification rate of the network and the corresponding changes in the output classification probability distribution, during different periods of the disorder (normal, pre-relapse, relapse). In order to prove the effectiveness of our approach for detecting relapses, we apply it to one of the largest datasets collected for biometrics in patients with psychotic disorders, with more than 18k days of collected data, and verify the output probability distribution change through extensive statistical analysis.**

Index Terms*— *relapse detection, person identification, psychotic disorder, biometrics, smartwatch

I. Introduction

Millions of people worldwide experience symptoms of psychotic disorders, with schizophrenia and bipolar disorder being the most common. Specifically, these diseases are classified as chronic diseases and are accompanied by repeated periods of relapse. Predicting psychotic relapses early, which could help patients live better, remain a major clinical issue that has not been solved yet. To this end, it could be useful to employ wearable consumer products that have enabled the reliable, unobtrusive, and remote collection of numerous behavioral and biometric signals through their sensors [1], [2]. This so-called "digital phenotyping" [3] has enabled significant advances in wearables for health purposes, leading to the fact that next-generation wearable technologies are about to transform hospital-centered healthcare practice into proactive and individualized care. Behavioral and biometric indexes have already been used in general medicine and sports, and nowadays, the evidence indicates that they could also be introduced into clinical psychiatry [4]. Using such signals to detect early indication of symptoms and prevent psychotic relapses is now one of the major research areas in psychiatry [5]–[7].

Some previous works have attempted to detect psychotic relapses from wearable/smartphone data. Valenza et al. [8] used Markov chains to model the mood in patients with bipolar disorder, using signals collected from a wearable shirt. Other methods have focused on a more statistics-based approach, such as [9], [10], where smartphone data were used to identify statistically significant anomalies in patients with schizophrenia.

Recently, a number of works have also successfully used biometric data from wearables in order to identify the identity

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

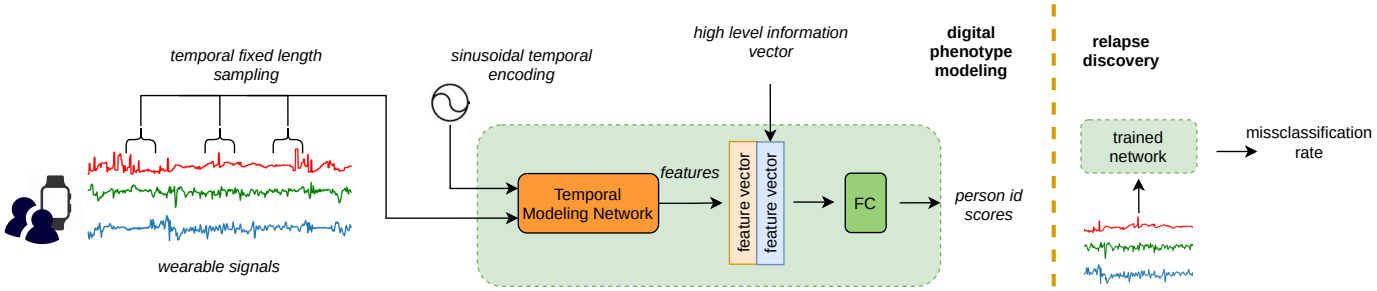

Fig. 1. The proposed framework for psychotic relapse detection via person identification. During training, our model learns the behavioral patterns of different wearers. Then, at inference, we look at the misclassification rate and the shift in the output score distribution at different periods of a psychotic disorder and detect relapses.

of the wearer [11], [12]. This result points to the fact that deep learning methods can now be successfully used to identify discriminating patterns of behavior. We argue that detecting the behavioral patterns that identify a person can be beneficial for detecting psychotic relapses since, during a psychotic relapse, the patient tends to adopt different behavioral patterns. This means that if we train a neural network to detect the behavioral patterns of a patient during remission, and then test it on the same patient during relapse, we should find changes in the distribution of the output probabilities of the network and, consequently, increased misclassification rate.

This paper presents a novel method for identifying relapses in patients with psychotic disorders using behavioral and cardiovascular signals obtained from long-term continuous wearable monitoring. Per our previous surmise, we first design a state-of-the-art architecture for digital phenotype classification (i.e., user identification). After that, we leverage the trained networks in order to show that they can successfully be used for discovering diverging behavior patterns when a user has transitioned to a psychotic relapse. Our contributions can be summarized as follows:

1) We introduce a novel method for detecting psychotic relapses by casting the problem of relapse detection as one of misclassification in neural networks trained for person identification. We also verify the changes in the output probability distribution scores for each person in different periods of a psychotic disorder (normal, pre-relapse, relapse), with thorough statistical analysis.
2) We design and build deep learning architectures for person identification using behavioral and cardiovascular signals from a wearable watch. We perform extensive exploratory studies regarding the architecture of the network, different input features, as well as different data augmentation strategies. The models achieve high classification rates, verifying the effectiveness of our models.
3) We train our models and validate our assumptions by detecting psychotic relapses on one of the largest ever collected datasets of biometric wearable data, including a total of 29 patients and $\sim 18,000$ days of recording data, spanning up to 2.5 years of continuous monitoring.

II. Person Identification from Digital Phenotypes

A. Framework Architecture

Our proposed methodology for detecting relapsing states in psychotic patients through person identification is outlined in Fig. 1. Our motivation for developing this framework stems from the fact that psychotic relapses are linked with considerable behavioral alterations. As a result, we intend to examine if this is reflected in the digital phenotype of patients and whether a trained network capable of identifying users based on phenotypes during remissions could efficiently identify users while confronting psychotic relapse or reveal abnormalities.

We first extract sequences of features from data collected continuously from wearable devices, aggregate them in a small temporal scale, and perform random sampling through their temporal dimension to reduce their dimensionality. The sampled feature sequence is then fed into a deep temporal modeling network architecture, which we train on the task of person identification, so that we can obtain the user's digital phenotype (the behavioral, circadian, and cardiovascular signals). We note that in this phase, training is carried out utilizing information from individuals with psychotic diseases who are currently in remission, while after training, we assess the trained network using data from the patient while they are in remission as well as during and right before a psychotic episode. Finally, we investigate how these behavioral shifts are reflected on the output distribution probability and classification accuracy of the network.

B. Training and Inference

The model input consists of a multidimensional signal $\mathbf{S}_t^L \in \mathbb{R}^N$, where L denotes the number of timesteps and N the extracted features, which corresponds to the daily recordings of a wearable user. Since the amount of the collected data varies throughout different days (see Section III-C), we randomly sample feature values from K input timesteps, resulting in a multidimensional signal $\mathbf{S}_t^K \in \mathbb{R}^N$ of fixed length K. Sampling is performed so that the resulting time series is temporally coherent; i.e., the temporal order of the samples in the original signal is retained. This technique is inspired by temporal segment networks [13] and allows us

to efficiently process signals of different lengths, which, as discussed in Section III-C, is a major obstacle in processing data collected via wearables. Also, it acts as a form of data augmentation, improving the ability of the resulting trained model to generalize. In addition, it helps model the long-range structure of the signals throughout the day and allows us to ignore redundant information in consecutive samples, helping avoid overfitting.

After temporal sampling, we augment the original signal $\mathbf{S}_t^K$ by concatenating it with two sinusoidal features, which denote the daily temporal cycle (similar to positional encoding [14]). Then, the fixed length sequences are fed into the temporal modeling network, which outputs an embedding vector of fixed size 512. The feature vector is now concatenated with a vector of additional high level features: the hours of the corresponding day that the user slept, and temporal weekly cycle information (day of the week). Finally, this augmented feature vector is fed into two fully connected layers, which output the identification scores; cross-entropy loss is used for training. After training the model and during inference, we perform the random sampling for each day five times and sum the unnormalized log scores in order to obtain the final score.

C. Temporal Modeling Network

For the temporal modeling network, we design two different architectures. One architecture is CNN-based and uses 1D convolutional kernels, while the second is based on Long Short-Term Memory (LSTM) layers. The CNN-based network consists of a fully connected layer followed by a ReLU activation and dropout. Then, five convolutional blocks follow, each consisting of a convolutional layer, ReLU, and dropout. The output of the last convolutional block is then fed into an adaptive average pooling layer in order to obtain a fixed feature vector of size 512. Finally, we use two consecutive fully connected layers (with ReLU in-between) on the feature vector to obtain the unnormalized log scores. In the LSTM-based architecture, the convolutional blocks are replaced with two bidirectional LSTM layers with dropout, and we get the output of the last time step as the feature vector. Note that, as we mentioned in Section II-B, the fixed feature vector output from the CNN or LSTM network is also concatenated with a vector that includes high-level information.

D. Data Augmentation

Training a person identification network, especially on limited amount of data, as in our case, may result to over-fitting phenomena. One extreme case of this was previously reported in [11], where unique signal patterns from the smartwatch sensors could lead to a system that recognizes the sensor (i.e., the smartwatch) rather than the actual person wearing it. Despite the fact that this behavior is more evident in raw signals, as in [11], similar "over-fitting" cases can also affect our study. To this end, we introduced an augmentation step, consisting of three distinct augmentation operations. Specifically, we considered:

- **Noise:** Perturbation of the data by sampling a normal distribution and applying this noisy signal to the initial signal in a multiplicative manner, as follows: $s_n = s(1 + r_n n)$, where s the initial signal, s_n its augmented counterpart, n is the noise signal, which is realized as a random variable sampled from $\mathcal{N}(0, 1)$, and r_n a user-defined hyper-parameter that controls the scale difference (used as 0.1 in our experiments). We used this multiplicative variant, instead of typical additive Gaussian noise, in order to preserve the scale of the initial signal and introduce only "high-frequency" noise that does not significantly affect the outline of the signal.
- **Random Mask:** We also used a dropout-inspired augmentation that has been proven effective to a wide range of deep learning applications, where we apply a random mask on the input signal. Since the input data are multidimensional (consisted of several features), this mask is a 2D sparse binary matrix that zeroes out specific input values in an unstructured way, i.e. in our experiments, the probability of masking each value is sampled from Bernoulli distribution with p = 0.05. This augmentation strategy pushes the network to learn useful representations even in the absence of particular features, thus increasing the generalization abilities of the network.
- **Mixup:** Finally, we also employed the state-of-the-art augmentation technique of mixup [15], where we merge two inputs x_1 and x_2 as $\lambda x_1 + (1 - \lambda)x_2$, where λ is sampled from a beta distribution, defined by $a = \beta = 0.2$. The augmented output is expected to be an interpolated version of the respective outputs y_1, y_2, with the same mixup factor λ. The simplicity of this approach, along with its notable regularization and generalization abilities, make it an ideal addition to our pipeline.

III. Database and Features Exploration

A. Database Collection

During the e-Prevention project, 38 people with psychotic spectrum illnesses were recruited and given a Samsung Gear S3 smartwatch by the University Mental Health Research Institute "Costas Stefanis". The participant signed an informed consent to disclose their data, which are anonymized in accordance with General Regulation(EU) 2016/679 [16]. The watch continuously monitored the user's linear and angular acceleration, heart rate, heart rate variability, sleeping schedule, and steps. Users were instructed to wear their smartwatches throughout the day, seven days a week, with the exception of charging and bathing. The charge time for the smartwatch was approximately two and a half hours, and daily charging is typically required. During charging, the data were uploaded and stored on a cloud-based platform [17].

Compared to previous projects, our database is among the largest [9], [18] ever recorded, spanning a maximum of 2.5 years of continuous patient monitoring, with more than 40 relapses recorded, both psychotic and non-psychotic. Relapses are common in psychotic diseases, and they play an important role in cognitive impairment, functional deterioration, and poor

TABLE I
STATISTICS OF THE RECORDED DATASET DURING E-PREVENTION [16]

	mean	std	sum
Days of participation	736.7	168.9	21,365
Days in database	613.0	187.71	17,777
Days in database (relapsed)	29.27	43.59	849.0
Hours per day	14.53	1.94	-

treatment response. A psychotic relapse is a clear clinical deterioration of a patient with the reoccurrence of psychotic symptoms (delusions, hallucinations) or manic, depressive, or mixed episodes with psychotic symptoms. The end of a relapse is defined as the moment when the symptoms lessen and the patient returns clinically and functionally to the state before the relapse. In terms of the pre-relapse period, there is no definite time when the patient approaches a psychotic relapse. In our study, the pre-relapse period is defined as four weeks before the annotated onset of the relapse, based on the clinicians' recommendations. This period's precise definition may help identify pre-relapse signs and forecast relapse.

Throughout the trial, our clinical staff conducted monthly follow-up examinations to check various psychiatric symptoms linked with their psychosis. At the start of data collection, patients were in a remission phase of their psychotic disorder. Relapses were recognized and examined using a monthly clinical examination of volunteer patients, changes in psychopathology as measured by the PANSS scale, and information gathered from treating psychiatrists, patients' families and carers, and clinic personnel. Relapses were noted by psychiatrists as mild, moderate, or severe based on their severity.

B. Experimental Dataset

According to our analysis of the collected data, the mean recording length for the entire dataset is approximately 14 hours per day, lower than the expected average of 18-20 hours per day in patients with good compliance. Among the causes for this are the following: a) ineffective photoplethysmography provided by the watch (for example, due to sweaty wrists), b) failure of the recording watch application, c) typical errors in use (not charging before the battery runs out, forgetting to wear), d) poor patient compliance, e) network issues. Nonetheless, the acquired data is of considerable magnitude and includes objective measurements of patients' everyday lives. Consequently, we will employ the recorded biosignals to examine the digital phenotypes of patients and their fluctuations during the course of their psychotic disorders.

After the initial collection of raw data, we extracted features using short-time analysis (five-minute interval windows). The preprocessing procedure is explained thoroughly in [16]. For our analysis here, we leverage a subset of the dataset consisting of data collected from 29 patients (2 with Schizoaffective disorder, 2 with Schizophreniform disorder, 15 with Schizophrenia, 8 with Bipolar I disorder and 2 with Bipolar II disorder), who have more than 180 days of recordings after data preprocessing. In our experimental dataset, eleven patients experienced 16 psychotic relapses over the study period.

C. Features Exploration

Our first goal is to find a concrete representation of each patient's digital phenotype in order to identify them using biosignals. In previous studies, we examined various features extracted from the raw signals and conducted extensive statistical analysis to reveal differences in physical activity and autonomic function patterns [16], [19]. Also, in [20], results indicate that heart rate measurements and RR intervals (time between successive heartbeats) correlate with patients' psychopathology changes. Considering the above and delving into the identification capabilities that the plethora of extracted features offer, we present an exploratory analysis to find an efficient representation to first use for the identification task.

Length of Data Sequence: First, we split the data into daily slots constituting the smallest possible measurement for detecting behavioral patterns by aggregating the five-minute feature values from 00:00 to 23:59. Therefore, if a patient wore the watch for the entire day, we would aggregate 288 values per feature (twelve five-minute intervals per hour). In Table I, the statistics for the recordings are accumulated. As noted previously, one of our study's greatest obstacles is that there is neither a steady amount of daily data collected for each user nor a fixed recording interval. This variation could have a negative impact on model training and introduce bias to the person identification task.To be more precise, if we use long sequences, we could remove users who only wear the watch for a few hours per day, while the network could learn to differentiate persons based on the recorded hours. Therefore, we prefer to create each data sequence by randomly choosing 5-minute interval vectors throughout the day, concatenating them, and sorting them by time. Therefore, a random representation of the day is generated. To determine our daily representation policy, we undertake trials with feature vectors of varying lengths and combinations of both network designs.

Features of Interest: We gather five groups of features that could prove useful in summarizing the users' phenotypes. We group the features as follows: a) heart rate variability statistics (the mean of RR intervals, their minimum and maximum values, as well as the Standard Deviation of RR intervals (SDRR)) denoted as *RR_intervals* in Fig. 2, b) heart rate (beats per minute) average statistics (the mean value, the minimum, the maximum, and the standard deviation of heart rate) denoted as *Heart Rate*, c) features of group (a), the Root Mean Square of Successive RR interval Differences (RMSSD) and number of recorded intervals, denoted as *RR_intervals + RR stats* , d) features of group (a) and the normalized powers of the high and low frequencies computed from the Lomb-Scargle power spectral analysis of RR intervals, as *RR_intervals + Lomb-Scargle feats*, and e) features of group (c) and the norm of the recorded linear acceleration. Note that we did not use angular velocity, due to the fact that its norm is highly correlated with the norm of the linear acceleration.

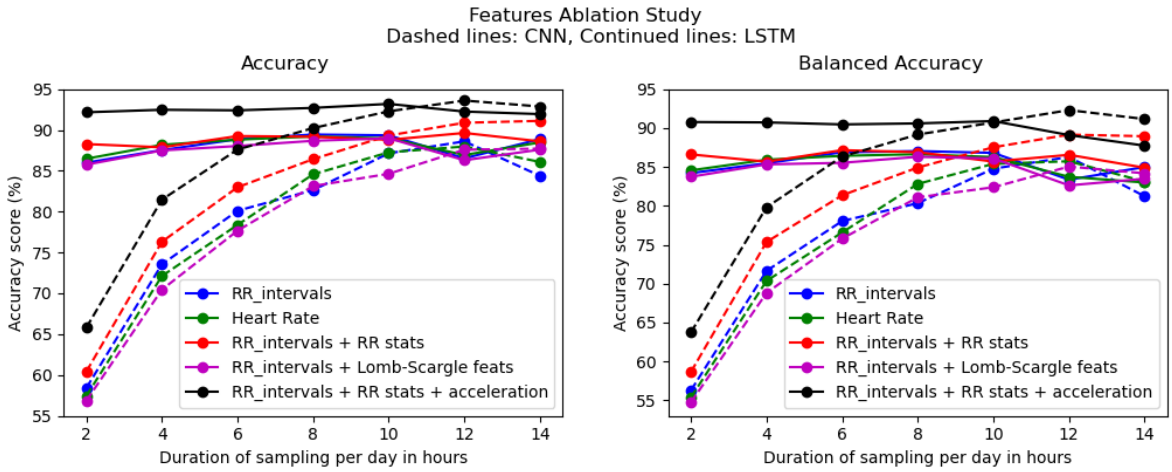

Fig. 2. Ablation study on different sets of features and length of time series for the classification task, reporting accuracy and balanced accuracy for both CNN (dashed lines) and LSTM (solid lines) architectures.

IV. Experimental Analysis

Experimental Setup: Experiments were evaluated on the collected dataset described in Section III-A. Training and testing were performed according to a five-fold cross-validation scheme, using only normal periods of all 29 patients, as expected. Both temporal modeling networks (CNN-based and LSTM) are trained with RAdam optimizer [21] for 100 epochs with an initial learning rate of 0.01, which drops to 0.001 at 75 epochs.

A. Feature Ablation

In the first ablation, we evaluate the performance of the proposed CNN and LSTM person identification networks for the different collected feature sets. In order to account for the unbalanced nature of patient data, we display in Fig. 2 results both in accuracy (on the left) and balanced accuracy (on the right) for the identification task of 29 patients. In the same Fig., we also investigate the effect of varying the sequence length. The x-axis represents the total duration of sampling per day in hours; for instance, 10 hours corresponds to a feature vector of 120 values length per feature (10 hours * 12 five-minute intervals per hour).

Feature Sets. Concerning the feature types, heart rate variability statistics perform slightly better than heart rate statistics. Adding RMSSD and the number of the recorded RR intervals, which could act as a certainty factor, we observe a small increase in the overall accuracy, indicating a slight improvement in the per-day classification. Finally, when we also add linear acceleration information, the model achieves the best results (in both balanced and unbalanced accuracy). Note that adding as input the normalized powers of the low and high frequency bands of the RR intervals did not result in increased performance. We also conclude that differences in network implementation have no substantial impact on the superiority of one set of characteristics over another, so for the rest of our experiments we proceed with the following set of features: *RR_intervals* + *RR stats* + *acceleration*.

Length of Data Sequence. From the same Fig. we also see that the length of the input sequence also affects the performance. We investigate a sample of daily recordings spanning from 24 to 168 timesteps (i.e., from 2 to 14 hours). We see that for the CNN architecture, random sampling at a sequence length of 144 (=12 hours) achieves the highest accuracy. For the LSTM implementation, a sequence length of 120 yields better results compared to the CNN, especially in terms of balanced accuracy. As a result, for the remainder of this study, we employ a multi-dimensional feature vector of 120 timesteps to represent each user's daily information. Another noteworthy conclusion we can draw is the difference in the number of samples (sequence length) required for the identification task between the two architectures. While the CNN requires information for more than eight hours to achieve high classification rates, LSTM networks, which are known for their efficiency with time series data, build powerful representations for a variety of sample sizes.

B. Architecture Ablation

Having explored different feature sets and discovered their contribution to phenotype identification, we now proceed to examine different architectures of the temporal modeling network and their ability to distinguish relapses. Specifically, we focus on comparing the two different temporal architectures, i.e. CNN vs LSTM, by training them on the data of all patients (29) under different augmentation schemes, namely scaling noise, random mask and mixup. Specifically, we add progressively each augmentation operation, forming an augmentation chain, in the order mentioned above. We also considered the architecture variations, where we included the temporal encoding information (see Section II-B), in order to understand if the information of the day cycle can be of importance. The results are summarized in Table II, where both accuracy and balanced accuracy - the arithmetic mean of sensitivity and specificity - are reported.

TABLE II

COMPARISON OF CNN AND LSTM ARCHITECTURES FOR THE PERSON IDENTIFICATION TASK DURING NORMAL PERIODS ACROSS ALL 29 PATIENTS. ACCURACY AND BALANCED ACCURACY IS REPORTED AFTER A 5-FOLD VALIDATION SCHEME FOR BOTH ARCHITECTURES (CNN & LSTM) OVER THE INITIAL EXTRACTED FEATURES, ALONG WITH TEMPORAL ENCONDING, AND THEIR AUGMENTED VERSION.

			Without Augmentation	**Adding Noise**	**+ Applying Random Mask**	**+ Mixup Samples**
CNN	**Base**	Acc.	**92.29**	89.58	89.19	70.36
		Bal.Acc.	**90.72**	88.42	87.58	65.41
	+ temporal encoding	Acc.	**91.56**	88.87	87.83	69.79
		Bal.Acc.	**89.81**	86.78	85.60	64.01
LSTM	**Base**	Acc.	**93.20**	91.32	92.89	91.54
		Bal.Acc.	**90.90**	88.64	90.46	89.18
	+ temporal encoding	Acc.	92.03	90.27	92.31	**92.57**
		Bal.Acc.	88.97	87.08	89.58	**90.06**

According to the reported results, the CNN models, as in the case of the previous ablation study, are much more sensitive than the LSTM variant. Specifically, adding noise to the CNN input considerably decreases the network performance. Contrary to such behavior, the LSTM network shows stable behavior across all settings. Regarding the extra temporal encoding information, we noticed a slight decrease in most settings. Despite this decrease, we consider this extra information important for the task at hand, since it can better model the behavior cycle within a day and we will continue its exploration in the upcoming evaluations. Note that the task of intermediate evaluation, namely person identification, is not perfectly aligned with the final task of relapse recognition and thus minor performance decrease when using such intuitive features is not considered a viable reason to drop them. Notably, the LSTM model, equipped with the temporal encoding and the full set of augmentations achieve similar performance to the top-performing system of base LSTM without any noise. To this end, we continue our experiments by assuming LSTMs, trained with the full augmentation pipeline, as the default option.

C. Distinguishing Different Periods of Psychotic Disorders

Given the well-performing LSTM system of the previous section, we now focus on the real problem: distinguishing the periods of psychotic disorders. Specifically, we want to explore if the information extracted from the person identification pipeline can be effectively used to detect a psychotic relapse (and ideally a pre-relapse) period following the concept of a behavioral change in the patient.

To do so, apart from the typical identification accuracy on the complete set of 29 patients during normal periods, we also reported the performance on the reduced dataset of eleven patients who faced relapses during the data collection. For this subset, we split the classification results in the three previously defined periods of a psychotic disorder: normal, pre-relapse, and relapse. These period-related results are presented in Table III. Reported accuracy results correspond only to balanced accuracy, since is more representative for the unbalanced nature of the relapse data. To better understand the detection ability of such a system, we also report the mean and median classification probability scores of the users in each period. This metric is more indicative of the model's outcomes because a biosignal could be classified on the right user but the probability of classification could be lower, for example, on the pre-relapse or the relapse period. In other words, even if the person is correctly identified, a consistent decrease in probability of detecting the correct person during a period may be a very useful indication of partially-changed behavioral patterns.

Table III, apart from providing information on the relapse detection problem, also contains an important ablation over the effect of a set of high-level characteristics, i.e., the sleep percentage and an encoding of the weekly cycle information, as mentioned in Section II-B. We consider the amount of sleep to be a crucial feature, because it fluctuates during relapse times, while the weekly cycle information may assist the network to distinguish patterns of the same day over different weeks.

Overall, we draw the following conclusions from the reported results in Table III: 1) Regarding the accuracy, the simplest version of the network (Base) with an extra feature including the hours of sleep achieves the highest results. However, when considering the set of 11 patients, we are mostly interested in detecting accuracy changes when comparing pre-relapse, relapse and normal periods. We observe that in general, the addition of high-level features widens the gap between normal and relapse phases in terms of accuracy. 2) The mean and median classification probability scores of the users in each period is indeed more indicative of the model's outcomes. 3) We conclude that all possible supplementary features (temporal encoding, sleep, and day of the week information) contribute to the detection of psychotic relapses because they intuitively capture a full profile of a person's daily activities.

D. Per-Person and Per-Period Identification Scores

To further understand the relapse detection ability, we now conduct a per-person specific study for each period (normal, pre-relapse, and relapse) and for each person, as presented in Table IV. Following its effectiveness form our previous analysis, we report the mean value of the per-person iden-

TABLE III

ABLATION STUDY OF EXTRA FEATURES (SLEEP AND DAY INFORMATION) FOR THE CASE OF LSTM ARCHITECTURE ON BOTH THE WHOLE COLLECTION OF 29 PATIENTS AND ON THE SUBSET OF 11 PATIENTS WHO FACED RELAPSES. BALANCED ACCURACY, MEAN AND MEDIAN PROBABILITY OF EACH EXPERIMENT AND PERIOD ARE REPORTED OVER THE THREE CONSIDERED PERIODS: NORMAL, PRE-RELAPSE AND RELAPSE.

	Metrics	**Balanced Acc. (29 patients)**	**Balanced Accuracy (11 patients)**			**Mean Probability (11 patients)**			**Median Probability (11 patients)**		
	Extra Features	**Normal**	**Normal**	**Pre-Rel.**	**Relapse**	**Normal**	**Pre-Rel.**	**Relapse**	**Normal**	**Pre-Rel.**	**Relapse**
Base	None	90.13	87.53	75.15	74.83	0.8685	0.8243	0.8227	0.9708	0.9096	0.9433
	Hours of sleep (HS)	**90.19**	**88.37**	77.44	72.52	0.8794	0.8215	0.8124	0.9873	0.9109	0.9472
	Day of Week (DW)	89.40	86.46	75.88	70.12	0.8632	0.8156	0.7756	0.9859	0.9332	0.8898
	HS + DW	89.76	87.65	76.88	70.35	0.8744	0.8076	0.7902	0.9909	0.8809	0.8980
+ Temporal encoding	None	89.04	85.93	75.02	71.53	0.8564	0.8288	0.8027	0.9465	0.9120	0.8947
	Hours of sleep (HS)	88.85	85.65	76.81	70.20	0.8579	0.8076	0.7812	0.9895	0.8832	0.8724
	Day of Week (DW)	88.07	85.16	75.20	68.84	0.8486	0.8032	0.7712	0.9687	0.8506	0.8791
	HS + DW	88.67	**86.00**	**74.28**	**69.86**	**0.8634**	**0.7954**	**0.7593**	**0.9690**	**0.8734**	**0.8498**

tification probability for each period. The reported results correspond to the LSTM model with all the additional extra information (temporal encoding, sleep information, day of the week) that achieved notable distinction between different periods in Table III.

In the Table IV, we also provide the absolute drop in mean identification probability for each user, for convenience. Our first observation is that we have a significant drop in the mean identification probability of the network, both during the pre-relapse and during the relapse periods. This shows that the network was relatively "confused" by the digital phenotype of the user when they went into relapse and right before it (pre-relapse). This motif of absolute drop can be seen in most users, with some exceptions, such as user #7, who had high identification probabilities throughout all his periods. Note that the availability of larger amount of data and mostly a larger set of patients could further help the discrimination between them and consequently the detection of different psychotic periods. In the next subsection we also include a statistical analysis of the per-patient scores.

The proposed rationale of phenotype identification can also be visually validated by Fig. 3, depicting the identification results for user #2 during a range of almost two and a half years. We can observe that, for the examined patient, the majority of the days during normal periods lead to correct identifications, while during relapses, a high percentage of miss-classifications occurs, indicating an anomaly.

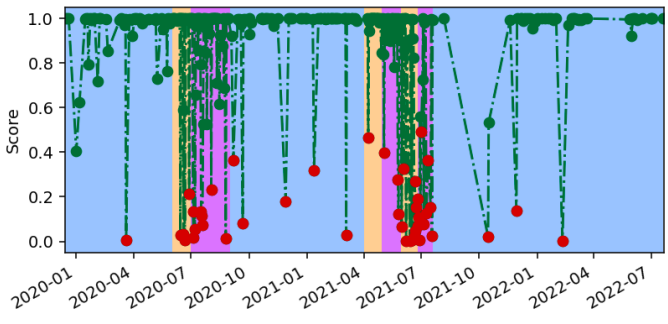

Fig. 3. Visualizing identification predictions of patient #2 : correct predictions are visualized with green color, while miss-classifications are denoted with red. The three normal periods are shaded with blue color, the three relapse periods with purple and the three pre-relapse periods with yellow.

E. Statistical Analysis of Scores

We also performed further statistical analysis on the outputs of our trained network. First, we collected for each patient across all folds the identification scores for each different phase and conducted pairwise single tail Mann-Whitney U-tests with Bonferroni correction. The results can also be seen in Table IV; each value in the absolute change section that is in bold denotes that scores during the phase at the left of the arrow (in the column title) were significantly greater than scores during the phase at the right of the arrow. While there are some cases (patients #7 and #11) where the detected changes are not significant, in the large majority of patients changes detected between the normal and pre-relapse and normal and relapse periods are significant and in the case of 3 patients, scores during the pre-relapse period are significantly greater than scores during the relapse period.

In addition, we also collected all the identification scores across all patients and folds. These amount to a total $N = 7942$ with 4012 scores for the normal periods, 1360 for pre-relapse, and 2570 for relapse periods. First, we show in Fig. 4 for each period the empirical cumulative probability distribution (eCDF). This curve shows for each value p in the x-axis the percentage of all scores that are lower than p. As it can readily be seen, identification scores during relapse periods of users tend to be lower than scores during the pre-relapse periods and normal periods (since at all points p of the curve, the percentage of relapse scores lower than p is larger compared to the percentage in pre-relapse and normals periods). The same holds when comparing the scores of the pre-relapse periods to scores during normal periods. The aforementioned conclusions are in-line with the analysis of the previous section.

Pairwise single tail Mann-Whitney U-tests with Bonferroni correction were also conducted for the overall scores and showed that a) the identification scores during the normal period were greater than scores obtained during the relapse period ($U = 4102847.5$, $p < 0.001$), b) scores during the normal period were greater than scores during the pre-relapse period ($U = 2413996.5$, $p < 0.001$), c) scores during the pre-relapse period were greater than scored during the relapse period ($U = 1578681$, $p < 0.001$).

TABLE IV

PER-USER AND PER PERIOD (NORMAL, PRE-RELAPSE, AND RELAPSE) MEAN IDENTIFICATION PROBABILITY SCORES. WE SHOW FOR CONVENIENCE THE ABSOLUTE CHANGE IN THE AVERAGE PROBABILITY BETWEEN ALL COMBINATION OF PERIOD TRANSITIONS. VALUES OF ABSOLUTE CHANGE ($x \rightarrow y$) IN BOLD DENOTE THAT THE SCORES DURING THE x PHASE WERE SIGNIFICANTLY (IN A STATISTICAL SENSE) GREATER THAN SCORES DURING THE y PHASE OF THE PATIENT. SEE TEXT FOR MORE DETAILS.

	Probability			Absolute Change		
ID	normal	pre-rel	relapse	normal → pre-rel	normal → relapse	pre-rel → relapse
1	0.971	0.944	0.826	**−0.027**	**−0.145**	−0.118
2	0.952	0.928	0.872	**−0.024**	**−0.080**	**−0.056**
3	0.963	0.913	0.997	**−0.051**	0.034	0.085
4	0.853	0.784	0.726	**−0.070**	**−0.127**	**−0.058**
5	0.905	0.718	0.765	**−0.187**	**−0.140**	0.047
6	0.905	1.000	0.794	0.095	**−0.112**	−0.206
7	0.966	0.983	0.974	0.017	0.008	−0.009
8	0.659	0.672	0.496	0.014	**−0.163**	**−0.176**
9	0.693	0.407	0.541	**−0.286**	**−0.152**	0.134
10	0.961	0.986	1.000	0.025	0.039	0.014
11	0.703	0.406	0.358	**−0.297**	−0.346	−0.049
all	0.866	0.795	0.759	−0.0723	−0.107	−0.035

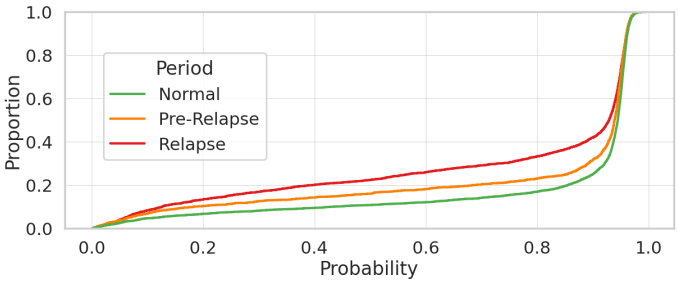

Fig. 4. Empirical cumulative probability distribution (eCDF) of identification scores during normal, pre-relapse, and relapse periods. As it can be seen, scores during the relapse and pre-relapse periods obtain lower scores more frequently, compared to normal periods.

F. Discussion and Limitations

Our work paves the way towards correlating psychotic disorders with phenotype identification, relying on the notion that psychotic relapses change the daily behavioral profile of the person and thus confuse an identification system. Our extensive experimental analysis verified this notion since we discovered significant changes in the output distribution scores of the networks during relapse and pre-relapse period, as well as a drop in the classification rate of the network. Further aspects of our framework should be examined more thoroughly to draw safe conclusions and assist clinicians. For example, the network architectures could be enhanced, and their effects on accuracy across the different periods fully demystified. Nonetheless, our work can stimulate further research in this direction.

V. CONCLUSIONS

We have introduced a novel framework for detecting relapses in patients with psychotic disorders. To that end, we have formulated the original problem as one of exploring misclassification rates and output probability score change of deep networks trained for identification of the digital phenotype of a specific user. The effectiveness of the proposed method has been validated on one of the largest datasets ever collected for biometrics in patients with psychotic disorders, where we found out that our assumption correctly holds and helps identify periods of psychotic relapse for specific users. We believe that our work opens up unexplored future opportunities for wearable-based health applications, through modeling of the digital phenotype.

REFERENCES

[1] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, "A review of wearable sensors and systems with application in rehabilitation," *J. Neuroengineering and Rehabilitation*, vol. 9, pp. 21, 2012.

[2] C. Boletsis, S. McCallum, and B. F. Landmark, "The use of smartwatches for health monitoring in home-based dementia care," in *Proc. Int'l Conf. Human Aspects IT Aged Population*, 2015.

[3] J. Torous, M. V. Kiang, J. Lorme, and J.-P. Onnela, "New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research," *JMIR Mental Health*, vol. 3, pp. e16, 2016.

[4] 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, pp. 603–609, 2017.

[5] D. Wiersma, F. J. Nienhuis, C. J. Slooff, and R. Giel, "Prodromes and precursors: Epidemiologic data for primary prevention of disorders with slow onset," *The American J. psychiatry*, vol. 152, pp. 967, 1995.

[6] N. Koutsouleris, C. Davatzikos, R. Bottlender, K. Patschurek-Kliche, J. Scheuerecker, P. Decker, C. Gaser, H.J. Möller, and E. M. Meisenzahl, "Early recognition and disease prediction in the at-risk mental states for psychosis using neurocognitive pattern classification," *Schizophrenia Bulletin*, vol. 38, pp. 1200–1215, 2011.

[7] P. McGorry, M. Keshavan, S. Goldstone, P. Amminger, K. Allott, M. Berk, S. Lavoie, C. Pantelis, A. Yung, S. Wood, and I. Hickie, "Biomarkers and clinical staging in psychiatry," *World Psychiatry*, vol. 13, pp. 211–223, 2014.

[8] G. Valenza, M. Nardelli, A. Lanat'a, C. Gentili, G. Bertschy, R. Paradiso, and Scilingo E. P., "Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis," *IEEE Jour. Of Biomedical and Health Informatics*, vol. 18, 2014.

[9] 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.

[10] P. Henson, R. D'Mello, A. Vaidyam, M. Keshavan, and J. Torous, "Anomaly detection to predict relapse risk in schizophrenia," *Translational psychiatry*, vol. 11, pp. 1–6, 2021.

[11] G. Retsinas, P. P. Filntisis, N. Efthymiou, E. Theodosis, A. Zlatintsi, and P. Maragos, "Person identification using deep convolutional neural networks on short-term signals from wearable sensors," in *Proc. Int'l Conf. ICASSP*, 2020.

[12] E. Maiorana, C. Romano, E. Schena, and C. Massaroni, "Biowish: Biometric recognition using wearable inertial sensors detecting heart activity," *arXiv preprint arXiv:2210.09843*, 2022.

[13] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. ECCV*, 2016.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[15] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*.

[16] A. Zlatintsi, P. P. Filntisis, C. Garoufis, N. Efthymiou, P. Maragos, A. Menychtas, I. Maglogiannis, P. Tsanakas, T. Sounapoglou, E. Kalisperakis, T. Karantinos, M. Lazaridi, V. Garyfalli, A. Mantas, L. Mantonakis, and N. Smyrnis, "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, pp. 7544, 2022.

[17] 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. AIAI*, 2020.

[18] M. Cella, Ł. Okruszek, M. Lawrence, V. Zarlenga, Z. He, and T. Wykes, "Using wearable technology to detect the autonomic signature of illness severity in schizophrenia," *Schizophrenia Research*, vol. 195, pp. 537–542, 2018.

[19] P. P. Filntisis, A. Zlatintsi, N. Efthymiou, E. Kalisperakis, T. Karantinos, M. Lazaridi, N. Smyrnis, and P. Maragos, "Identifying differences in physical activity and autonomic function patterns between psychotic patients and controls over a long period of continuous monitoring using wearable sensors," *arXiv preprint arXiv:2011.02285*, 2020.

[20] E. Kalisperakis, T. Karantinos, M. Lazaridi, V. Garyfalli, P. P. Filntisis, A. Zlatintsi, N. Efthymiou, A. Mantas, L. Mantonakis, T. Mougiakos, I. Maglogiannis, P. Tsanakas, P. Maragos, and N. Smyrnis, "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.

[21] L. Liu, P. Jiang, H.and He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.