A comparative study of autoencoder architectures for mental health analysis using wearable sensors data

M. Panagiotou, A. Zlatintsi, P. P. Filntisis, A. J. Roumeliotis, N. Efthymiou, and P. Maragos

School of ECE, National Technical University of Athens, Greece
{el16703, aroumeliot}@mail.ntua.gr, {filby, nefthymiou}@central.ntua.gr,
{nzlat, maragos}@cs.ntua.gr

Outline

Introduction

- This study is an ongoing work of the **e-Prevention project** (`http://eprevention.gr`), targeting to innovative e-Health services for patients' effective monitoring.
- The project leverages digital phenotyping to collect data of patients with psychotic disorders from smartwatch.
- An approach that could be used for relapse detection in patients suffering from psychotic disorders is **the sensor-based anomaly detection**.

Data Collection

- 24 patients with psychotic disorders.
- From Samsung Gear S3 smartwatch that continuously monitored acceleration, angular velocity and heart rate.
- Using the Tizen API provided by the smartwatch, we collected information about the sleep schedule and steps.
- Continuous recordings 24/7 (except 2 hours during charging).
- Clinicians annotates the patient's condition as either stable or relapsing. Also denoting the specific period of the relapse and its severity marked as low, moderate or severe.

I. Maglogiannis, A. Zlatintsi, A. Menychtas, D. Papadimatos, 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

Data Preprocessing I

Extracted sequences of features found to contain significant information as shown in a previous study. The following features are extracted:

- Mean energy of the accelerometer.
- Gyroscope norm.
- Mean heart rate.
- RR interval of heart rate variability.
- Mean frequency in the LF and HF bands of the heart rate.
- The value of the width of the ellipse (SD1) in the Poincare recurrence plot.
- The percentage of correctly identified pulses in the given interval.
- The sine and cosine representations of corresponding seconds to model the chronological order of the time-series.

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*, vol. abs/2011.02285, 2020

Data Preprocessing II

- 5-mins intervals.
- Filling Method: Median.
- The data of each patient are considered as a multivariate time-series $\mathbf{X}_{L\times d}$, where L denotes the total length (varying for each patient) and d the number of features (which is 10). Then, we apply an I-length rolling window with stride 1, creating a total of $N = L - I + 1$ subsequences, thus, resulting in a $M_{N\times I\times d}$ tensor for the data of each patient.
- We split the data in subsequences of 24 hours examining this way the patients' daily patterns.
- A total of 10 patients had sufficient data after preprocessing.

Figure: Number of days used in our experiments after preprocessing.

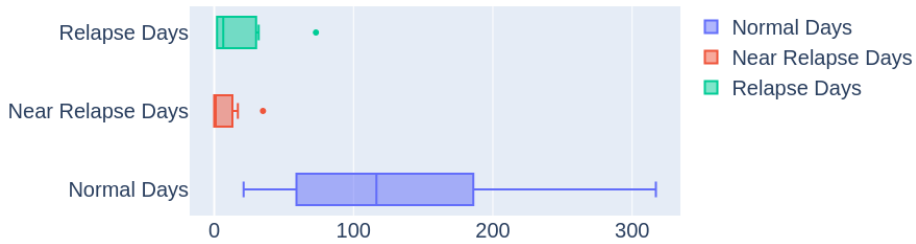
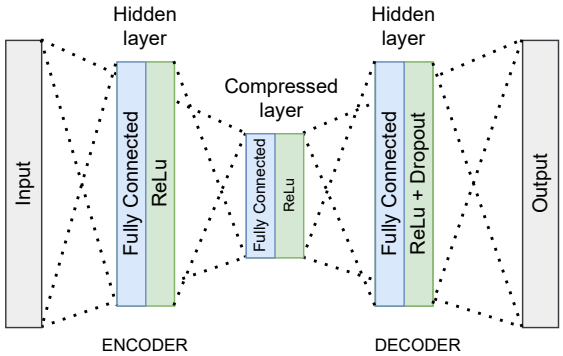


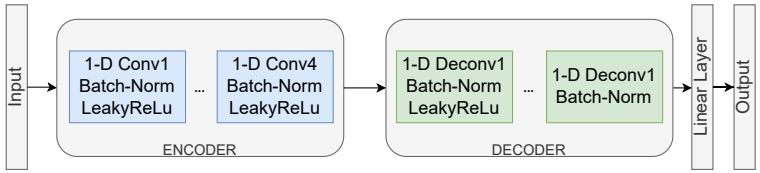
Table: Demographics information

Male/Female	6/4
Age (years)	30.60 ± 7.31
Education (years)	13.8 ± 1.99
Illness dur. (years)	7.3 ± 7.06

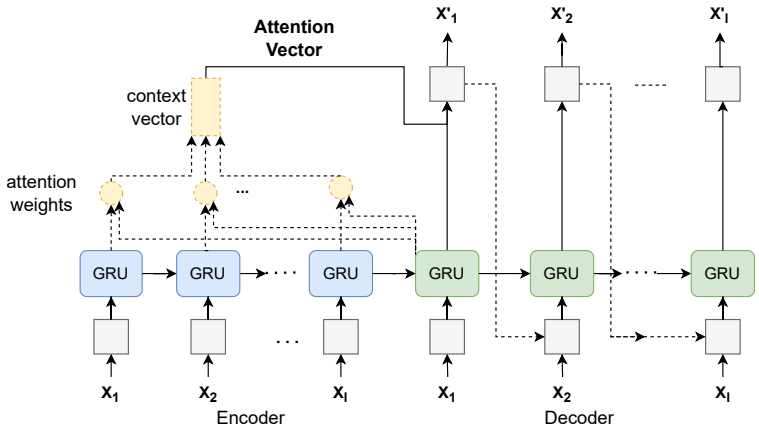
Autoencoder Architectures



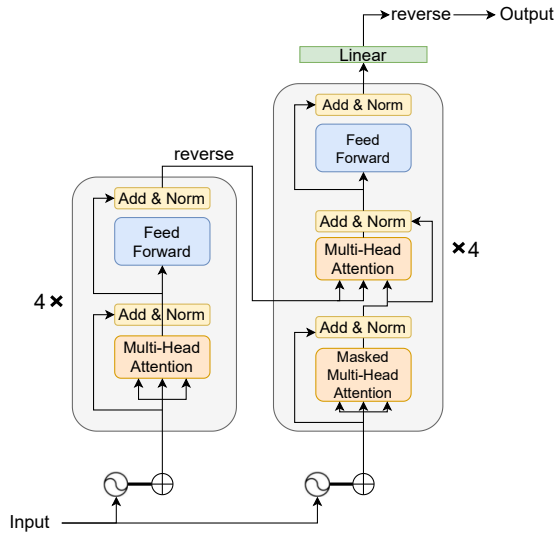
(a) FNN AE architecture.



(b) CNN AE architecture.



(c) GRU with attention AE architecture.



(d) Transformer architecture.

Training & Evaluation of Anomalies I

1. Splitting the data into Normal, Near Relapse and Relapse periods.
2. Separate the normal data in three sets, i.e., train, validation and test set, with a split of $60 - 20 - 20(\%)$, respectively.
3. Normalize the training data in the $[0, 1]$ range and the values in the other data sets are transformed appropriately.
4. Train each individual architecture using only data corresponding to "normal" periods.
5. Evaluate it to "unseen" normal and relapse data.

Training & Evaluation of Anomalies II

How do we detect relapse?

1. Calculate the reconstruction error vector with size $l \times d$ of each point i given by $e^{(i)} = |x^{(i)} - x'^{(i)}|$, originated by the trained model between the predictions $x'^{(i)}$ and given data $x^{(i)}$ in the validation set. The error vectors $e^{(i)}$ for the points in the sequences are used to compute the mean (μ) and covariance (Σ) of a multivariate normal distribution that is the expected error distribution.
2. Extract the Mahalanobis distance referred to as the "anomaly score" between the predicted points in the test set and the Gaussian distribution that calculated in the validation set, as follows:

$$a^{(i)} = \sqrt{(e^{(i)} - \mu)^T \Sigma^{-1} (e^{(i)} - \mu)} \tag{1}$$

Anomaly Condition

if $a^{(i)} \geq$ *Threshold* **then**
    Anomaly
else
    Normal
end if

- Evaluation under multiple thresholds using Receiver Operating Characteristic Area Under Curve (ROC AUC) and Precision-Recall Area Under Curve (PR AUC).

P. Malhotra *et al.*, "Long short term memory networks for anomaly detection in time series," in *Proc. ESANN*, 2015

Training & Evaluation of Anomalies III

- Personalized Scheme.
- Global Scheme.
  - Globally: evaluated to all patient.
  - Individually: evaluated individually, per patient.

Results & Discussion I

Personilized Scheme

- Best performing model is the CNN AE with PR and ROC AUC scores, 76% and 61%, respectively.
- Patient #1 (11 moderate relapse days) has the best performance in the Transformer model with PR and ROC AUC scores, 97% and 97%, respectively.
- All results surpass Random Classifier (baseline) results.

Table: Results for PR-AUC (personalized scheme).

Patients	FNN	CNN	Transformer	GRU	Random
#1	0.94	0.95	**0.97**	0.91	0.91
#2	**0.05**	0.04	0.02	0.03	0.03
#3	**0.54**	0.46	0.43	0.44	0.53
#4	0.26	**0.34**	0.18	0.19	0.18
#5	**0.63**	0.57	0.60	0.61	0.63
#6	0.70	**0.72**	0.63	0.67	0.68
#7	0.82	0.86	**0.87**	0.85	0.86
#8	0.83	**0.87**	0.65	0.81	0.85
#9	0.79	**0.80**	0.45	0.75	0.68
#10	0.97	0.95	0.94	**0.97**	0.97
Median	0.75	**0.76**	0.61	0.71	0.68

Table: Results for ROC-AUC (personalized scheme).

Patients	FNN	CNN	Transformer	GRU	Random
#1	0.94	0.96	**0.97**	0.93	0.91
#2	**0.49**	0.40	0.22	0.36	0.28
#3	**0.57**	0.53	0.49	0.49	0.52
#4	**0.39**	0.39	0.35	0.29	0.22
#5	0.44	0.28	**0.45**	0.40	0.42
#6	**0.49**	0.49	0.39	0.42	0.48
#7	0.56	**0.69**	0.69	0.64	0.62
#8	0.72	**0.78**	0.64	0.60	0.72
#9	**0.78**	0.75	0.28	0.58	0.42
#10	0.91	0.88	0.81	**0.94**	0.91
Median	0.57	**0.61**	0.47	0.54	0.50

Results & Discussion II

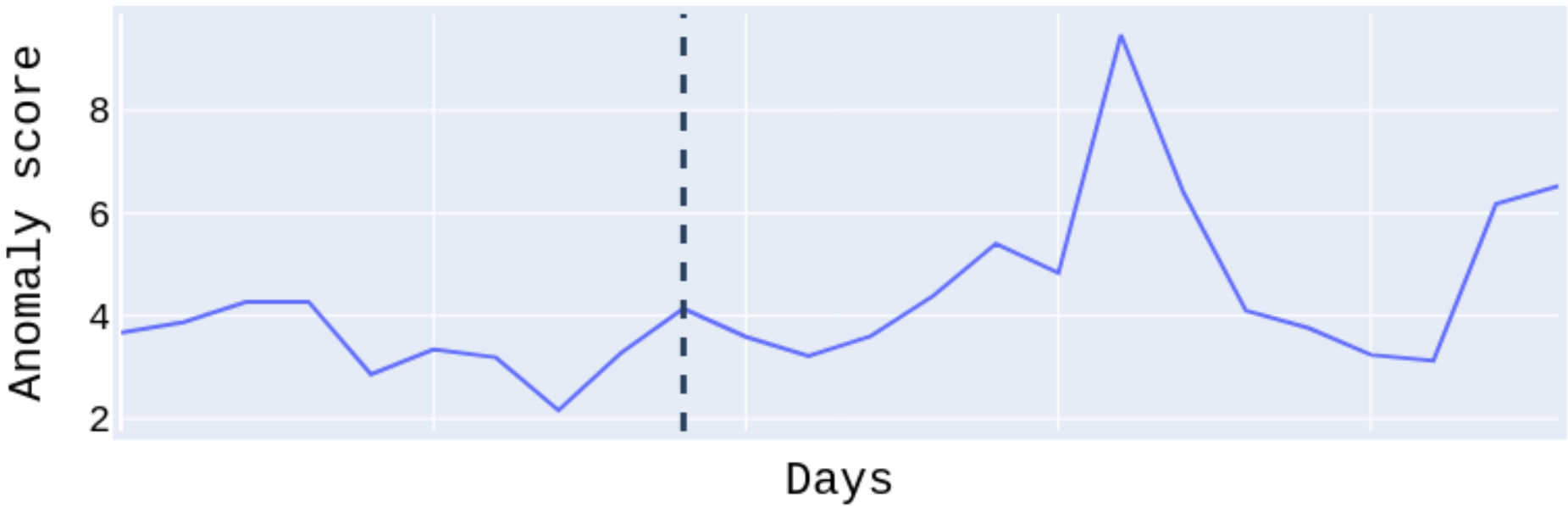


Figure: Anomaly score of Patient #1

Results & Discussion III

Global Scheme

- Best performing model is the FNN AE, which was evaluated individually, with PR and ROC AUC scores (77%) and (62%), respectively.

Table: Results for PR and ROC AUC (Global Scheme (Global) and Global scheme evaluated individually (Median)).

	FNN	**CNN**	**Transformer**	**GRU**	Random
PR AUC					
Median	**0.77**	0.71	0.76	0.73	0.68
Global	0.48	0.49	0.47	**0.52**	0.50
ROC AUC					
Median	**0.62**	0.58	0.52	0.57	0.50
Global	0.47	0.51	0.45	**0.53**	0.50

Results & Discussion IV

- The difference of the personalized CNN AE and the global FNN AE model that was evaluated individually is relatively **small**.
- Better performance for #1 and #10 patients, with moderate and severe relapse, respectively. Lowest performance for #2 patient, with 2 days of low severity relapse.
- **t-tests** contacted, shown statistical significance for $6/10$ patients, with p-value < 0.05.

Severity Levels impact in anomaly detection

Table: Reconstruction error of the best performing models for Low and Moderate relapses.

	CNN AE		FNN AE	
Patients	**Low**	**Moderate**	**Low**	**Moderate**
#5	0.007824	0.006617	0.004493	0.004342
#6	0.005098	**0.005481**	0.003960	0.003824
#7	0.006257	**0.006515**	0.005915	**0.005951**

Table: Reconstruction error of the best performing models (global scheme) for Low, Moderate and Severe relapses.

Patients	**FNN AE**	**CNN AE**
Low	0.002598	0.003650
Moderate	0.002629	0.003759
Severe	**0.002796**	**0.004076**

Conclusions

- Best performance of Personalized scheme: CNN AE model.
- Best performance of Global scheme evaluated individually: FNN AE model.
- Notice statistical significance between our architectures and the random classifier.
- Severity level of relapses: the more severe a relapse is, the easier it is to detect.
- Possible **future directions** include the addition of more patients, experimentation with other and possible more informative feature representations and investigation of possible differentiations between wakefulness and sleep.

Thank you for your attention!