AUGMENTING TRANSFORMER AUTOENCODERS WITH PHENOTYPE CLASSIFICATION FOR ROBUST DETECTION OF PSYCHOTIC RELAPSES

N. Efthymiou[1], *G. Retsinas*[1], *P. P. Filntisis*[1], *and P. Maragos*[1,2]

[1]Institute of Robotics, Athena Research and Innovation Center, Maroussi 15125, Greece

[2]School of ECE, National Technical University of Athens, 15773 Athens, Greece

{nefthymiou, george.retsinas, pfilntisis }@athenarc.gr, {maragos}@cs.ntua.gr

ABSTRACT

Recently, deep autoencoder architectures have received attention for the problem of unsupervised anomaly detection. Detecting psychotic relapses in mental health patients is a crucial challenge, often framed as anomaly detection, given the limited availability of data during relapsing states. In this paper, motivated by the fact that during relapses patients tend to undergo behavioral changes, we augment the classical autoencoder architecture with extra patient identification components. We show that formulating the problem as one of both signal reconstruction and patient identification largely improves the overall precision and robustness of relapse detection and significantly outperforms previous methods with a relative improvement of 15%. In addition, we also explore multiple ways to fuse the identification and reconstruction errors into a unified anomaly score that outperforms the results achieved by each error in isolation.

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

1. INTRODUCTION

Nowadays, a notable trend has emerged within the mental health community. Practitioners, clinicians, and engineers are increasingly collaborating to harness the power of digital phenotyping [1, 2]. By tapping into the extensive data provided by wearables and smart devices, these professionals aim to offer a more nuanced evaluation of mental health, with an ultimate goal of predicting relapses [3]. Wearables offer the advantage of being unintrusive, blending into daily life and yielding authentic data without observation biases. Their continuous "in situ" recording not only captures obvious behavioral changes but also highlights subtle, often overlooked shifts, that potentially signal an impending relapse [4].

Typically, relapse detection can be tackled both as a supervised problem (when enough data are available in both relapse and non-relapse states) and as unsupervised (when limited data is available for relapsing states). Supervised learning approaches for correlating the appearance of relapses with physiological data have mostly focused on either statistical significance testing or classification of hand-crafted features using traditional machine learning algorithms. Consequently, a variety of feature representations have been proposed in such medical settings using data from wearables [5, 6]. Unsupervised approaches for relapse detection using autoencoders have been presented in [7, 8, 9], while in [10], clustering models were used. A different paradigm was employed by [11] where relapse detection was cast as a miss-classification problem by networks trained to predict the identity of the users from their biosignals, while in [12] a self-supervised method was proposed using survival analysis. An important milestone towards unsupervised relapse detection was the organization of the e-Prevention Challenge I in ICASSP 2023 which focused on relapse detection in patients with psychotic disorders. The winning method of [13] constituted of an ensemble of deep autoencoder architectures (based on CNNs, LSTMs, and Transformers), trained in a personalized scheme (one model per patient) and using the reconstruction error as an anomaly score. Hamieh et al. [14] presented also a simpler shallow autoencoder for tackling the problem. Finally, [15] presented a method based on isolation forest that was trained on carefully handcrafted features.

In this paper, we present a novel framework that augments the traditional autoencoder architectures that are used for anomaly detection. More specifically, we propose to seamlessly integrating autoencoder-based relapse detection with misidentification methods [11, 5]. This integration can be used during inference to create a robust anomaly detection score. Overall, our contributions are summarized as follows:

1. We augment classical autoencoder architectures for anomaly detection with extra patient identification components. Specifically, we adopt a common "universal" architecture for all patients. The joint objective of reconstruction and patient identification results in substantial improvement in detecting relapses, even without taking into account the identification predictions of the network.
2. We leverage the learned feature representations and train patient-specific models to construct an *identification anomaly error* that further improves relapse detection scores compared to the classical *reconstruction anomaly error*.
3. Finally, we explore the fusion of the *identification anomaly*

This research work was supported by the project "Applied Research for Autonomous Robotic Systems" (MIS5200632) which is implemented within the framework of the National Recovery and Resilience Plan (NNRP) "Greece 2.0" (Measure: 16618- Basic and Applied Research) and is funded by the European Union-NextGenerationEU.

Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024), Seoul, Korea, April 2024

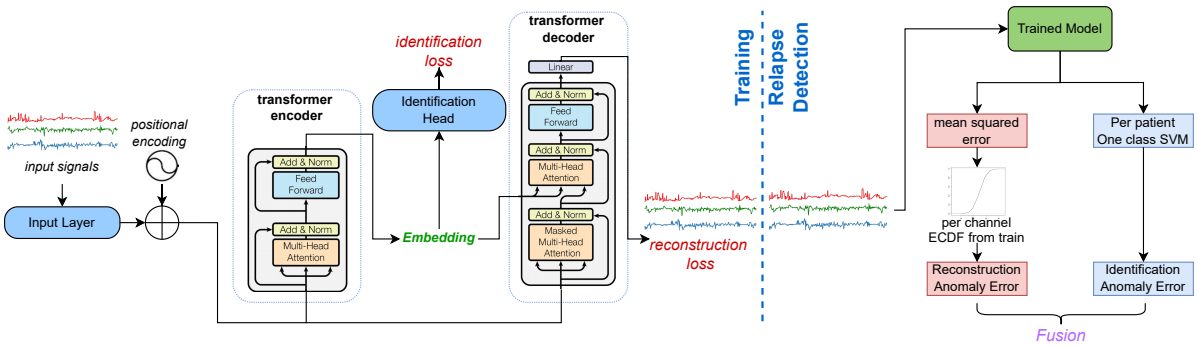

Fig. 1. Framework overview. A transformer autoencoder is first trained with reconstruction and identification losses. At inference, the final anomaly score is calculated using the reconstruction error ECDF and the identification score from personalized one-class SVM models.

error and *reconstruction anomaly error* into a single *joint anomaly error* that achieves significantly higher performance compared to each error in isolation.

2. DATASET

We use the dataset of the Track 2 of the ICASSP 2023 e-Prevention Grand Challenge I [16] which includes smartwatch biosignals from 10 patients with psychotic disorders, monitored over the course of 6 months, all of whom relapsed at some time (once or more). User recordings are split into separate days and the data for each day include continuous signals of *linear acceleration* (from the accelerometer), *angular velocity* (from the gyroscope), *heart-rate* and *RR-interval* (from photoplethysmography - PPG). The values of the signals were aggregated over 5 seconds to mitigate the effect of each individual sensor noise in classification tasks, as pointed out in [5]. The final training set of the dataset includes data acquired only while the patients were stable (1906 days), while the validation (533 days) and test sets (544 days) span both stable and relapsing periods.

3. DETECTION OF PSYCHOTIC RELAPSES

Here we describe the motivation and architecture of our proposed framework for psychotic relapse detection. In a nutshell, we borrow from two diverse, yet effective, approaches, namely relapse detection as anomaly prediction [13] and relapse detection as person miss-identification [11] and propose a seamless way to combine these into a single efficient framework. This combination is multifaceted: we include both schemes to the training phase, with two separate losses forming a multi-task loss, as well as during evaluation, with an explored range of possible fusion approaches.

3.1. Framework Architecture

We base our method on the framework of [13]. First an autoencoder-based architecture is trained on data during normal (non-relapse) periods. After training, a Cumulative Distribution Function (CDF) is computed for the reconstruction error on the training data. Finally, at inference time the value of the CDF is used as an anomaly score.

Architecture-wise, the underlying model is a transformer-based autoencoder the detailed architecture of which is depicted in Fig. 2. Here, the input signals are projected to an embedding space and a positional encoding is used to capture the temporal correlations. The Encoder and Decoder sub-modules are multi-layered and multi-headed transformers. Finally, the output of the decoder is projected into the same size as the input. The experimental section includes an ablation over the transformer's hyper-parameters.

Note that in the initial work [13], multiple families of neural networks were considered, including CNNs, where for each patient the best model was selected via the validation set. In contrast, in this work, we focus only on the transformer architecture, aiming to showcasing its prominence, while avoiding the more costly solution of multiple networks, one for each person. *Instead, we propose a "universal" model capable to predict anomalies across different persons.* Our motivation is two-fold: 1) cross-patient patterns for anomaly detection - we may extrapolate useful information for a possible relapse from other patients and 2) storage efficiency - a single model should be trained and stored for all patients.

3.2. Introducing Phenotype Classification

We proceed to integrate the person identification concept in the existing framework, following the success of our previous works [5, 11]. To do this, we train the network to identify patients from the encoder's embedding (where all information is distilled) using an additional identification head. We also add the corresponding cross-entropy loss as an auxiliary loss term during the framework's training (Fig. 2).

In total, the autoencoder is now trained with the following joint criterion: $\mathcal{L}(\mathbf{x}, y) = \mathcal{L}_{MSE}(d(e(\mathbf{x})), \mathbf{x}) + \lambda \mathcal{L}_{CE}(c(e(\mathbf{x})), y)$, where $\mathbf{x}$ is the input signal, y the person's identity, while e, d and c are functions, implemented as neural networks, that represent the encoder, the decoder and the classification head, respectively. Moreover, $\mathcal{L}_{MSE}$ is the mean squared error loss, $\mathcal{L}_{CE}$ the cross entropy loss and λ a hyper-parameter that balances the contribution of the latter.

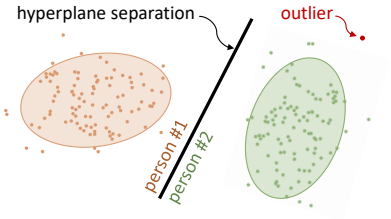

Fig. 2. Visual example of the inability of classification to capture outliers as anomaly. Here, the depicted hyperplane separates the two classes. When an outlier is introduced, which may indicate a sudden change in behavior, it is classified with high probability to be indeed person #2. Nonetheless, unless a distribution model is considered, this rationale cannot capture potential critical outliers.

Our key motivation behind this extra classification criterion lies on the fact that under psychotic relapses patients undergo considerable behavioral alterations which can "trigger" a miss-classification of the person's identity. Apart from this end-goal aspect, the joint training of signal reconstruction and person classification could potentially enhance both tasks, highlighting their correlation and further implicitly promote the generation of highly effective encoder embeddings.

3.3. Inference: Which Criterion to Use?

The system formulation of two simultaneous tasks leads to the following questions: "which criterion to use?" and "how can we combine both the task-related metrics?".

First, since relapse-detection is done per-day, to generate the corresponding anomaly scores we first extract 10-feature sequences from 5-minute data slices and combine them into four-hour windows. We use all of the available four-hour duration windows in overlapped spans of three hours to provide a robust representation.

Following the anomaly prediction rationale, the Reconstruction error, processed by a CDF, is used as an anomaly score. On the other hand, following the phenotype classification rationale, a straightforward way to define an anomaly score is to directly use the predicted probability p of the corresponding user for each sample (in fact, $1 - p$). In simple terms, as the probability drops, there is a stronger indication of a possible anomaly. However, preliminary experimentation showed that this metric led to under-performance and was not a good indicator for behavioral changes.

This under-performance can be attributed to the classification space, where classes are separated by hyper-planes, as depicted in Fig. 2. Under this assumption, the introduction of an outlier, in the sense that the new prediction is not close to an existing one, can be faithfully classified to a specific class with a high probability, following this hyper-plane separation rationale. On the other hand, if we model the distribution of each class, after having them trained with the aforementioned classification pipeline, we can potentially detect anomalies which correspond to behavioral changes. To this end, we proposed the following procedure: 1) we train for each user a one-class SVM on features extracted from the transformer encoder on his samples. 2) Subsequently, we use for each sample in the validation test the score of the one class svm as a *identification anomaly error*. This pipeline acts as a lightweight personalized post-processing, applied on the features of the "universal" model. As we will show this method significantly boosted relapse detection results.

Model	Size	Heads	Layers	AUROC	AUPRC	Mean
Base	32	8	2	**0.6254**	0.6343	**0.6299**
	32	8	4	0.6206	0.6285	0.6246
	32	16	4	0.6212	**0.6368**	0.6290
	64	8	2	0.6224	0.6319	0.6271
	128	8	2	0.6221	0.6296	0.6259
Augmented with Identification Loss	32	8	2	**0.6469**	**0.6565**	**0.6517**
	32	8	4	0.6223	0.6312	0.6267
	32	16	4	0.6205	0.6300	0.6252
	64	8	2	0.6290	0.6456	0.6373
	128	8	2	0.6217	0.6305	0.6261

Table 1. Ablation study on various transformer architectures. The evaluation is displayed on the validation set for the Mean Square Reconstruction Error.

Score Fusion: The proposed framework results to two distinct anomaly scores. To further promote the effectiveness of our method, we consider different techniques for fusing the *reconstruction anomaly score* and the *identification anomaly score*: 1) non-linear combination (i.e., product), 2) fitting a linear model on the two scores, and 3) fitting a linear model on the per-input-signal reconstruction anomaly score, combined with the identification score.

4. EXPERIMENTAL ANALYSIS

Experimental Setup: Our models are trained and evaluated using as input sequences of 10 features extracted from a slice of 5 minutes data. These features are the mean norm of linear and angular acceleration, the mean of RRintervals and heart rate, the major axis of the Poincare ellipse, the normalized low and high powers of the Lomb-Scargle periodogram, temporal enconding of the recording time, and percentage of valid sample in the 5 minutes data. During testing, because every day has a different number of windows, we use the mean anomaly score for the final per-day anomaly score. We trained each scheme three times for 80 epochs on the training dataset of non-relapse days and used the validation set to select the best model. As relapse detection metrics we use the Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), and their harmonic mean (averaged over the three runs).

Transformer Architecture Ablation: First, we examined the impact of a single universal model with and without the proposed identification loss, under different architectural choices (encoder/decoder hidden size, #heads, #layers). Evaluation is performed using only the reconstruction error. Table 1 presents this architectural ablation for both training variations (with and without the extra loss). Notably, identification loss, only as a training option, provides a consistent

	Best Val. Reconstruction Error			Best Val. Identification Error			Best Val. Combination		
	Validation								
	AUROC	AUPRC	Mean	AUROC	AUPRC	Mean	AUROC	AUPRC	Mean
Reconstruction	0.6299	0.6360	0.6329	0.5486	0.5636	0.5562	0.6033	0.6139	0.6086
Identification Error	0.5945	0.6307	0.6125	0.6581	0.6882	0.6731	0.6407	0.6752	0.6579
Combination	0.6491	0.6940	0.6716	0.6539	0.6895	0.6717	**0.6670**	**0.7129**	**0.6900**
	Test								
Reconstruction Error	0.6592	0.6631	0.6611	0.5689	0.5889	0.5789	0.6314	0.6401	0.6357
Identification Error	0.5744	0.6139	0.5941	0.6430	0.6725	0.6578	0.6335	0.6595	0.6465
Combination	0.6598	0.7095	0.6847	0.6629	0.6940	0.6785	**0.6865**	**0.7332**	**0.7099**

Table 2. Relapse detection metrics (rows) when we select as final anomaly score: 1) the reconstruction error, 2) the identification error, 3) their combination. Each super-column shows according to which criterion the final model was selected. For all metrics, higher is better.

	Validation			**Test**		
λ	AUROC	AUPRC	Mean	AUROC	AUPRC	Mean
1.0	0.6670	0.7129	0.6900	0.6865	0.7332	0.7099
0.8	0.6816	0.7087	0.6951	0.7033	0.7395	0.7214
0.6	0.6923	0.7010	0.6966	0.7002	0.7158	0.7080
0.4	**0.7017**	**0.7389**	**0.7203**	**0.7231**	**0.7667**	**0.7449**
0.2	0.6678	0.6781	0.6730	0.6934	0.7113	0.7023
0.1	0.6628	0.6948	0.6788	0.6768	0.7161	0.6964

Table 3. Ablation study (combination score) on the scale λ of the identification loss during training for the anomaly score of the combination.

Global	Pers.	Global-Signal	Pers.-Signal
0.7291	0.7758	0.7241	**0.7812**

Table 4. Mean AUROC-AUPRC when doing supervised fusion of anomaly scores using ridge regression.

increase in relapse prediction. The best overall model (32 hidden size / 8 #heads / 2 # layers) exceeds the rest base models and achieves results comparable to the winning method of the e-Prevention Challenge [13].

Evaluation Strategies & Metrics: Next we explore the performance of our identification anomaly score on relapse detection, as well as the formulation of a joint anomaly score. Table 2 presents relapse detection results based on three different anomaly scores: 1) *reconstruction error*, 2) *identification error*, and 3) their *non-linear combination*. Importantly, during the training phase, we select the model with the best detection metric. Nonetheless, in this setup we have three distinct scores to be used for calculating the detection metric. Thus it is not fair to select the best validation model according to a specific score and then evaluate our system using another scoring option. To this end, depending on which criterion is used, three different models were selected from the validation set and evaluated against all three scores. As it was expected, the model selection affects significantly the final prediction score. We observe that the use of the proposed identification error results in higher performance, compared to using the reconstruction error. In addition, for both the validation and test set, using a criterion based on the combination further improves the result in almost every case.

Moreover, for the sake of a thorough exploration, Table 3 presents an ablation study for the λ hyper-parameter that scales the $\mathcal{L}_{CE}$ cross entropy loss and as a result, controls the contribution of the identification part of the framework. In this experiment, the combination fusion technique is selected. It should be noted that previous experiments have a predefined scale of $\lambda = 1$.

Methods		AUROC	AUPRC	Mean
Unsupervised	Avramidis et al.[15]	0.5839	0.6263	0.6051
	Hamieh et al.[14]	0.6072	0.6347	0.6209
	Calcagno et al. [13]	0.6469	0.6509	0.6489
	Combination (Ours)	**0.7231**	**0.7667**	**0.7449**
Supervised/Fusion	Pers.-Signal (Ours)	**0.7343**	**0.8291**	**0.7812**

Table 5. Final results on the e-Prevention Challenge I dataset.

Supervised Fusion: As a final exploratory experiment, we explored a supervised fusion of the anomaly scores, by running ridge regression of positive values on the output anomaly scores of the validation set and the corresponding labels, and then testing. We considered four different methods: 1) one global linear model, 2) personalized per-person linear models, 3) one global linear model which takes into account the per-signal reconstruction error (instead of the mean reconstruction error as used till now), and 4) personalized per-person models with per-signal reconstruction error weights. Our results, shown in Table 4 show that best fusion of scores is achieved when taking into account the per-signal reconstruction error and using personalized linear models for each patient. These findings suggest the potential significance of individual signals in detecting relapses, prompting the need for further investigation towards this direction in future works.

Final Results: Finally, in Table 5 we show comparison with the other SoTA methods on the e-Prevention Challenge I dataset. As we can see, our developed framework achieves a relative improvement of almost 15% in the unsupervised setting and 20% when using our supervised fusion.

5. CONCLUSION

We have presented a novel framework for relapse detection that integrates traditional autoencoder architectures and person identification. Our experiments showed that our dual approach, focusing jointly on data reconstruction and patient identification, significantly and consistently outperformed previous methods by a large margin. This paves the way for more tailored and timely interventions, highlighting the evolving role of technology in patient monitoring and care.

6. REFERENCES

[1] J.P. Onnela, "Opportunities and challenges in the collection and analysis of digital phenotyping data," *Neuropsychopharmacology*, vol. 46, no. 1, pp. 45–54, 2021.

[2] Y. Liang, X. Zheng, and D. D. Zeng, "A survey on big data-driven digital phenotyping of mental health," *Information Fusion*, vol. 52, pp. 290–307, 2019.

[3] A. Cohen, J. A Naslund, S. Chang, S. Nagendra, A.t Bhan, A. Rozatkar, J. Thirthalli, et al., "Relapse prediction in schizophrenia with smartphone digital phenotyping during covid-19: a prospective, three-site, two-country, longitudinal study," *Schizophrenia*, vol. 9, no. 1, pp. 6, 2023.

[4] E. Rodriguez-Villa, U. M. Mehta, J. Naslund, D. Tugnawat, S. Gupta, J. Thirtalli, A. Bhan, et al., "Smartphone health assessment for relapse prevention (SHARP): a digital solution toward global mental health," *BJPsych Open*, vol. 7, no. 1, pp. e29, 2021.

[5] 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.

[6] 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.

[7] M. Panagiotou, A. Zlatintsi, P. P. Filntisis, A. J. Roumeliotis, N. Efthymiou, and P. Maragos, "A comparative study of autoencoder architectures for mental health analysis using wearable sensors data," in *Proc. of the 30th European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, 2022.

[8] D. A. Adler, D. Ben-Zeev, V. WS 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, 2020.

[9] K. Wang, Y. Zhao, Q. Xiong, M. Fan, G. Sun, L. Ma, and T. Liu, "Research on healthy anomaly detection model based on deep learning from multiple time-series physiological signals," *Scientific Programming*, vol. 2016, 2016.

[10] J. Zhou, B. Lamichhane, D. Ben-Zeev, A. Campbell, A. Sano, et al., "Predicting psychotic relapse in schizophrenia with mobile sensor data: Routine cluster analysis," *JMIR mHealth and uHealth*, vol. 10, no. 4, pp. e31006, 2022.

[11] N. Efthymiou, G. Retsinas, P. P. Filntisis, C. Garoufis, A. Zlatintsi, E. Kalisperakis, V. Garyfalli, T. Karantinos, M. Lazaridi, N. Smyrnis, and P. Maragos, "From digital phenotype identification to detection of psychotic relapses," in *Proc. IEEE Int'l Conf. on Healthcare Informatics (ICHI-2023)*, Houston, TX, USA, June, 2023.

[12] E. Fekas, A. Zlatintsi, P.P. Filntisis, C. Garoufis, N. Efthymiou, and P. Maragos, "Relapse prediction from long-term wearable data using self-supervised learning and survival analysis," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, June*, 2023.

[13] S. Calcagno, R. Mineo, D. Giordano, and C. Spampinato, "Ensemble and personalized transformer models for subject identification and relapse detection in e-prevention challenge," in *Proc. Int'l Conf on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, June*, 2023.

[14] S. Hamieh, V. Heiries, H. Al Osman, and C. Godin, "Relapse detection in patients with psychotic disorders using unsupervised learning on smartwatch signals," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, June*, 2023.

[15] K. Avramidis, K. Adsul, D. Bose, and S. Narayanan, "Signal processing grand challenge 2023–e-prevention: Sleep behavior as an indicator of relapses in psychotic patients," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, June*, 2023.

[16] "https://robotics.ntua.gr/eprevention-sp-challenge/".