# Alert Mechanism for Relapse in Bipolar Disorder Using Deep Learning

Katerina Menychta
University of Piraeus, Department of Digital Systems
*Big Data and Analytics*
*AM: me2421*

***Abstract*—This paper presents a hybrid anomaly detection framework for identifying relapse episodes in patients with bipolar disorder using behavioral data from the ePrevention dataset. The methodology involves training an Autoencoder exclusively on samples from a single patient, in order to learn typical behavioral patterns. During inference, reconstruction losses from the Autoencoder serve as anomaly scores. A threshold-based alert mechanism is then employed: if the reconstruction error exceeds a predefined threshold by a significant margin or if multiple consecutive daily alerts occur, a relapse is considered highly probable based on the model's output.**

**In contrast to conventional multi-patient approaches, this implementation adopts a personalized modeling strategy, treating the selected patient as a case study. The trained Autoencoder was trained on days with no relapse (label = 0), but applied to the entire dataset, including relapse samples (label = 1), and the results showed a clear increase in anomaly scores during known relapse periods, based on which the labels were assigned to the dataset. Although the current approach is limited to a single patient due to resource and time constraints, it provides a promising foundation for developing individualized early warning systems in mental health monitoring. Future work should aim to incorporate data from multiple patients to improve the generalization of the model.**

## I. Introduction

Bipolar disorder is a psychiatric condition characterized by alternating episodes of mania and depression. Early identification of potential relapses is critical. Recent research has explored machine learning techniques to address this.

Recent research emphasizes the growing potential of machine learning and passive sensing for forecasting mental health conditions. In [1], Recurrent Neural Networks (RNNs) were used to model temporal behavioral data derived from smartphone usage, demonstrating how sequence-based models can capture latent patterns associated with depressive episodes. This work illustrates the applicability of deep learning in real-world, noisy behavioral datasets collected from mobile devices.

The work of [2] further reinforces this direction by showcasing the predictive power of physiological and behavioral features—specifically sleep and activity data—in detecting early signs of relapse in bipolar disorder. This study used multimodal monitoring and highlighted that subtle behavioral changes can precede clinical deterioration, a finding directly relevant to anomaly detection frameworks.

Furthermore, the importance of ensemble strategies and interpretability is supported by [4], where Random Forests were shown to provide both accuracy and feature importance—an asset when used in combination with unsupervised models. Although more complex architectures like Long Short-Term Memory (LSTM) networks [5] and attention-based Transformer models offer improved temporal modeling capabilities, they come at the cost of reduced explainability and higher computational demands.

Lastly, [6] argues that in highly imbalanced classification tasks—such as relapse detection—precision-recall plots are more informative than ROC curves, a recommendation that influenced our evaluation strategy.

Collectively, these studies underscore the importance of personalized, interpretable, and efficient models for mental health monitoring. Our proposed hybrid framework builds upon these insights, combining unsupervised anomaly detection with supervised classification and advocating for future integration of temporal modeling and multimodal signals.

This work proposes the use of an Autoencoder trained on non-relapse data to detect behavioral deviations, using reconstruction loss as an anomaly score to trigger relapse alerts.

## II. Methodology

*Implementation Details:* All preprocessing and modeling steps were implemented in Python 3.8. The key libraries used included `pandas` and `numpy` for data manipulation, `scrubkit-learn` for scaling and utility functions, and `TensorFlow` (v2.12) with `Keras` for constructing and training the autoencoder model.

The final autoencoder architecture consisted of a feedforward neural network with an input layer matching the number of daily features adding Gaussian noise, a single hidden layer (encoding dimension = 10), and a reconstruction output layer with `tanh` activation. The model was compiled using the Adam optimizer and trained with a Mean Absolute Error (MAE) loss function for robustness to outliers. Early stopping was employed to prevent overfitting, using a patience of 5 epochs to prevent overfitting.

All scripts were executed in Jupyter Notebook, and the complete codebase, including preprocessing scripts and evaluation notebooks, is available on GitHub for reproducibility.

The proposed methodology for relapse detection using behavioral data consists of the following steps:

- **Preprocessing and filtering:** The dataset is first filtered to include only samples labeled as $label = 0$, representing normal behavior. These are used exclusively for training the autoencoder. All features are then standardized to have zero mean and unit variance using the StandardScaler.
- **Autoencoder training:** A feedforward autoencoder is trained in an unsupervised manner on the normalized $label = 0$ samples. The architecture includes a reduced latent dimension (bottleneck) to force compact feature representation. Mean Absolute Error (MAE) is used as the loss function to improve robustness to outliers during training.
- **Reconstruction error computation:** After training, the autoencoder is used to reconstruct all samples in the dataset. The reconstruction error for each sample is calculated using the MAE between the input and the reconstructed output, serving as an anomaly score.
- **Threshold-based anomaly detection:** A statistical threshold is defined based on the distribution of reconstruction errors for normal data (e.g., the 95th percentile). Samples exceeding this threshold are flagged as potential relapses.
- **Temporal validation and interpretation:** The anomaly detection results are analyzed in the context of known relapse periods. Metrics such as precision, recall, and false positive rate are computed. Reconstruction error distributions and time-series plots are used to interpret model performance and identify patterns.

## III. Results

*Data Preprocessing and Dataset Construction:* The original dataset consisted of anonymized '.zip' files containing granular physiological and behavioral data collected from multiple patients over time. Each zip archive included raw sensor readings, such as heart rate, accelerometry, gyroscope data, sleep events, and step counts, often recorded multiple times per day. The raw data was heterogeneous, partially incomplete, and organized per patient and per day, requiring a structured preprocessing pipeline to extract useful features and labels for model training.

For this study, we focused on a single patient (ID: `5f615ea99e38890013062039`) with clearly annotated relapse periods. This decision was motivated by the need to perform personalized modeling and by the imbalance of relapse data across the cohort. The final dataset preparation followed a multi-step process, as outlined below:

- **Initial data inspection:** Scripts such as `stats_output.py` and `tar_files_missing.py` were used to analyze the availability and completeness of `.zip` files across the study period, and to determine the data coverage for the selected patient.
- **Label assignment:** The `assign_labels.py` script assigned a binary label (0 = normal, 1 = relapse) to each day based on known relapse episodes. For patient 5031, relapse labels were assigned to the following periods: 2020–09–30 to 2020–10–29 (severe episode), and 2021–03–13 to 2021–03–18 (moderate episode).
- **Feature extraction:** The `create_final_dataset.py` script parsed each zip file and extracted key daily-level features. These included heart rate statistics (mean, max), gyroscope variance, linear acceleration, step count, calories burned, and sleep duration. All extracted features were aggregated at the daily level.
- **Profiling and quality checks:** The generated dataset was profiled using `dataset_profile.py`, which summarized distributions, identified inconsistencies, and flagged missing values.
- **Data cleaning and engineering:** Final cleaning was performed using `data_preprocessing.py`, which imputed missing values (where applicable), removed low-quality samples, and constructed derived features (like zero imputation flags) where appropriate. After this step, the dataset contained well-aligned daily observations with consistent feature formats and corresponding labels.

Importantly, this preprocessing step enabled the construction of a structured and clean dataset for training and evaluation. Furthermore, due to a strong imbalance in label distribution (many more normal days than relapse days), the model training focused exclusively on $label = 0$ samples, making an unsupervised autoencoder an appropriate fit for the task and translating the task into an anomaly detection one.

*Final Dataset Description and Preprocessing Pipeline:* The final dataset used in this study was derived from a larger, anonymized collection of daily behavioral and physiological measurements collected via wearable devices as stated previously.

The transformation from raw multi-patient sensor data to a clean, single-patient, daily-level dataset involved several key preprocessing steps:

- **Patient filtering and labeling:** The study centered on patient 5031, who had documented relapse periods. Labels were assigned based on date ranges of known severe and moderate episodes, with $label = 1$ representing relapse days and $label = 0$ indicating normal behavior.
- **Daily feature aggregation:** For each available date, the data from the high-frequency sensor were summarized in a single daily record. The resulting features captured core behavioral and physiological indicators, including:
  - *Heart rate:* `hr_mean`, `hr_max`, `hr_median`, `hr_valid_count`
  - *Gyroscope:* `gyr_var`, `gyr_max`, `gyr_energy`
  - *Linear acceleration:* `linacc_var`, `linacc_max`, `linacc_energy`
  - *Mobility:* `steps_walking`, `steps_running`, `total_distance`, `total_calories`
  - *Sleep:* `sleep_ratio`
- **Missing and invalid data handling:** Several sensor-derived features had missing or zero values. Based on domain knowledge and documentation from similar wear-

able studies, values such as zero heart rate or gyroscope variance were considered biologically implausible and treated as invalid. The following corrections were applied:
  - Zero or negative values were replaced with `NaN` and then imputed using the median of the valid values (for most columns).
  - For `gyr_max`, missing values were imputed using synthetic values sampled from a truncated normal distribution based on the existing feature's mean and standard deviation.
  - Outlier capping was applied to `hr_max`, with values above 216 (a known physiological threshold) clipped to that maximum.
  - For `hr_valid_count`, entries with zero count were preserved only for relapse-labeled rows; for normal days ($label = 0$), such rows were dropped.
- **Feature tracking:** For transparency and future interpretability, indicator flags such as `<feature>_was_zero` or `<feature>_was_invalid` were introduced during preprocessing to trace which values had been imputed or flagged as originally missing.
- **Column selection:** After cleaning, columns deemed redundant, low informative, or inconsistent were removed. Specifically, `hr_min`, `hr_valid_count`, and intermediary flag columns like `linacc_max_was_zero` were excluded from the final modeling dataset.

The resulting dataset, saved as `daily_features_cleaned.csv`, contains one record per day for the selected patient, with 18 cleaned and engineered features and a binary label. Given the strong imbalance between normal and relapse-labeled samples, the modeling focused solely on the abundant $label = 0$ data during training. This decision not only improved the stability and generalization of the model but also aligned with the anomaly detection paradigm of learning from normal behavior and identifying deviations.

*Personalization vs Generalization:* This study adopts a personalized modeling approach by focusing on a single patient diagnosed with bipolar I disorder (patient ID: `5f615ea99e38890013062039`). The autoencoder is trained exclusively on this individual's non-relapse ($label = 0$) data, enabling the model to learn patient-specific behavioral patterns without interference from inter-subject variability.

This design choice was motivated by both methodological clarity and data availability. Since behavioral baselines and physiological responses can vary widely between individuals, patient-level modeling helps improve anomaly sensitivity and reduce false positives. Additionally, data from other patients either lacked sufficient relapse annotations or exhibited inconsistent coverage, limiting their utility for generalized modeling.

While personalization improves precision for the selected subject, it limits the direct generalizability of the model to broader populations. Future work may explore extending this approach to multi-patient settings using techniques such as domain adaptation, patient clustering, or federated learning.

*Model Selection and Experimental Considerations:* The choice of model architecture and training strategy was guided by both empirical experimentation and insights from recent literature on anomaly detection in behavioral and health monitoring contexts. Given the objective of identifying deviations in individual daily patterns over time, an unsupervised autoencoder was selected for its ability to learn a compact representation of normal behavioral signals without requiring labeled anomaly data. Additionally, the dataset exhibited a strong imbalance between the two classes, with a significantly higher number of $label = 0$ (non-relapse) samples compared to $label = 1$ (relapse). This further justified the use of an unsupervised learning approach, as it enabled the model to focus on accurately capturing the structure of normal behavior and detecting deviations, without relying on scarce labeled relapse data.

Initially, alternative architectures such as Long Short-Term Memory (LSTM) networks and Transformer-based models were considered. However, they were not adopted due to the following reasons:

- **Personalized modeling:** Since the approach focuses on learning a baseline of one patient's normal behavior, a static feedforward autoencoder was sufficient to capture per-day deviations without modeling temporal dependencies across days.
- **Computational simplicity:** Autoencoders offer faster training and fewer hyperparameters than recurrent or attention-based models, making them more practical for iterative experimentation.
- **Greater explainability:** The reconstruction error from autoencoders can be decomposed per feature, allowing the identification of which behavioral dimensions (e.g., steps, heart rate) contributed most to the anomaly—something not easily achievable with complex temporal models like LSTMs or Transformers.

Several experimental variations were also explored to improve generalization and anomaly separation:

- **Dropout regularization:** Dropout layers were introduced to prevent overfitting and encourage redundancy in the latent representation. However, this degraded performance by disrupting the model's ability to reconstruct important input patterns consistently.
- **Denoising input noise:** A Gaussian noise layer was tested to simulate a denoising autoencoder setup, which is commonly used in biomedical anomaly detection. Despite promising results in literature, this method reduced anomaly sensitivity in our dataset, likely due to the subtle nature of relapse signals being masked by noise injection.

Both techniques were informed by prior work (e.g., [Ali et al., 2024], [Zhou et al., 2020]) that highlight the value of robustness-oriented architectures in health anomaly detection. Nevertheless, empirical results on this specific dataset showed that these regularizations diminished detection accuracy.

As a result, the final model design retains the core structure of a feedforward autoencoder with a small encoding dimension (to limit memorization) and Mean Absolute Error (MAE) loss (to handle outliers robustly). This configuration was found to offer the best balance between simplicity, interpretability, and detection effectiveness.

*Reconstruction Loss Statistics:* To better understand how the autoencoder's reconstruction loss differs between normal and relapse days, we computed descriptive statistics per label. Table I summarizes the count, mean, standard deviation, minimum, quartiles, and maximum of reconstruction losses for both groups.

TABLE I
RECONSTRUCTION LOSS STATISTICS BY LABEL

| Label | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| 0 (Normal) | 25 | 0.1141 | 0.1452 | 0.0026 | 0.0161 | 0.0490 | 0.1595 | 0.4631 |
| 1 (Relapse) | 25 | 0.6126 | 1.0654 | 0.0030 | 0.0204 | 0.3757 | 0.5227 | 4.9936 |

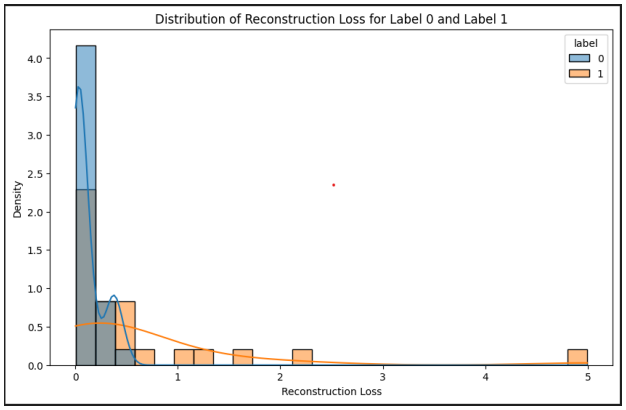


Fig. 1. Overlayed distribution of reconstruction loss for label 0 (normal) and label 1 (relapse) samples.

*a) Final Observations:* Figure 1 provides a combined view of the reconstruction loss distributions for both classes. It clearly shows that the majority of $label = 0$ samples cluster tightly near zero, while $label = 1$ (relapse) samples exhibit a broader and heavier-tailed distribution.

This further supports the anomaly detection strategy used in our work: the autoencoder generalizes well to normal behavior and fails more noticeably on relapse days, resulting in significantly higher reconstruction errors. The separation between these distributions justifies the use of percentile-based thresholds for alert generation and validates the effectiveness of the model's learned representation of normal behavior.

*b) Interpretation:* The statistics indicate a clear separation between the two classes:

- **Normalized distribution:** The mean reconstruction loss for relapse days ( 0.613) is over five times higher than that for normal days ( 0.114), demonstrating that the autoencoder reconstructs normal behavior much more accurately.
- **Greater variance:** The standard deviation among relapse days (1.065) is also much larger than for normal days (0.145), indicating relapse behavior produces a wider range of reconstruction errors.
- **Overlap in lower tails:** Both groups share similar losses in the minimum and 25th percentile (around 0.002–0.020), implying that some relapse days were reconstructed similarly to normal days, possibly due to mild relapse symptoms.
- **High-end separation:** The maximum and 75th percentile values differ significantly—normal days cap at 0.463, while relapse days reach up to 4.994—supporting the use of a percentile-based threshold (e.g., 95th percentile on normals) to distinguish relapse signals.

Computing these statistics was essential to justify the thresholding strategy and confirms that reconstruction loss serves as an effective anomaly score: relapse-day losses cluster away from the normal-day distribution, making the detection task feasible and meaningful.

*Threshold Selection Justification:* To convert reconstruction losses into binary relapse alerts, a threshold-based decision rule was applied. Specifically, any daily sample with reconstruction error exceeding a given percentile of the non-relapse ($label = 0$) loss distribution was flagged as an anomaly.

Two candidate thresholds were evaluated: the 95th and the 98th percentile. The choice impacts the model's sensitivity (recall) and specificity (false positive rate). As shown in our results, the 95th percentile threshold (0.3774) detected 48% of relapse days (12 out of 25), with a false positive rate of 8% (2 out of 25 normal days misclassified). In contrast, the stricter 98th percentile threshold (0.4220) reduced the false positive rate to 4%, but at the cost of sensitivity—only 36% of relapse days (9 out of 25) were detected.

Given the clinical importance of early relapse detection, we prioritized higher recall to avoid missing potential relapse signals, even at the cost of slightly more false positives. Consequently, the 95th percentile was selected as the operational threshold, striking a better balance between sensitivity and specificity for the studied patient.

Figure 2 and Figure 3 illustrate the distribution of reconstruction losses for both relapse ($label = 1$) and normal ($label = 0$) samples under the 98th and 95th percentile thresholds, respectively.

In both cases, relapse samples tend to exhibit higher reconstruction loss values, but the 95th percentile threshold captures more of the tail of the relapse distribution. This leads to better separation between the two classes, as reflected in higher recall. While the 98th percentile provides a stricter cut-off, it fails to capture a portion of moderate relapse days whose reconstruction losses lie just above the 95th percentile line but below the 98th. Therefore, visual inspection of the density plots supports the quantitative findings and further motivates the use of the 95th percentile threshold in our final model.

*Relapse Period Evaluation (Case Study):* To assess the real-world applicability of the model, we evaluated its performance during a known relapse episode from September 30, 2020, to

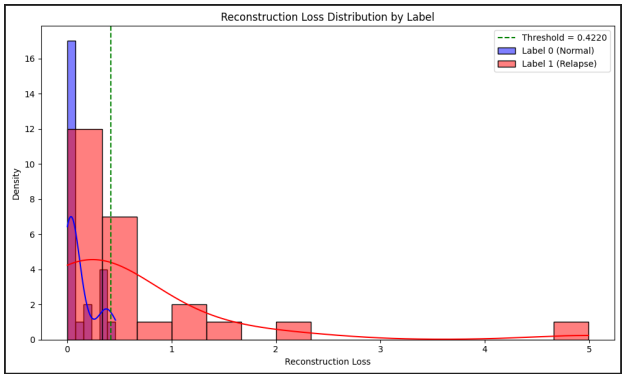


Fig. 2. Reconstruction loss distribution with threshold at 98th percentile (= 0.4220).

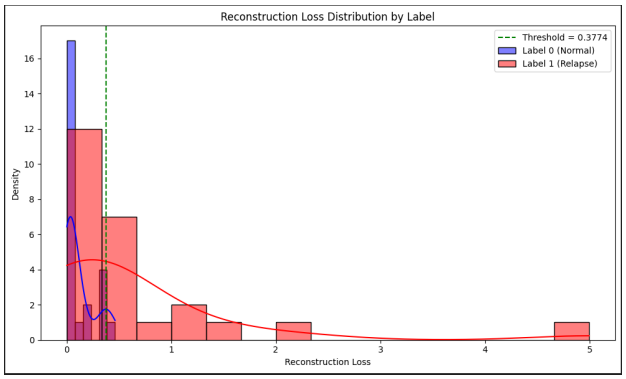


Fig. 3. Reconstruction loss distribution with threshold at 95th percentile (= 0.3774).

October 29, 2020. This period was previously annotated as a severe episode for the selected patient (see methodology).

Using the threshold determined from the 95th percentile of reconstruction losses for normal samples ($\theta = 0.3774$), we observed the following:

- **Label 1 (Relapse) samples:** 11 days occurred within the relapse period.
  - **True positives:** 7 of these days had reconstruction losses above the threshold, yielding a detection rate of **63.64%**.
  - **Mean reconstruction loss:** 0.7583
  - **Median reconstruction loss:** 0.4563
- **Label 0 (Normal) samples:** 16 days in the same period were labeled as normal.
  - **False positives:** 2 days exceeded the anomaly threshold, giving a false positive rate of **12.50%**.

These results demonstrate that the model successfully flagged the majority of relapse days within the target period, while maintaining a relatively low false alarm rate among normal days. The higher mean and median reconstruction losses for relapse samples reinforce the model's ability to capture behavioral deviations during clinical deterioration.

*Feature Attribution and Global Error Patterns:* To interpret which features contributed most to the autoencoder's anomaly alerts during relapse periods, we conducted both local and global attribution analyses based on reconstruction error.

*c) Local Attribution for Relapse Anomalies:* Using the scaled test data, we selected the 11 relapse-day samples from the annotated severe episode (2020–09–30 to 2020–10–29) that exceeded the anomaly detection threshold. For each of these days, we computed the squared reconstruction error between the input and the autoencoder's output across all features.

The top contributing features varied across days, but certain physiological indicators appeared consistently:

- On **2020–09–30**, the highest loss contribution came from `hr_mean`, followed by `hr_max` and `hr_median`, suggesting abnormal cardiovascular patterns compared to the patient's normal baseline.
- On **2020–10–02**, the dominant features included `hr_mean`, `total_calories`, and `total_distance`, highlighting disrupted physical activity or energy expenditure.

To identify common patterns across all detected anomalies, we averaged the per-feature squared errors across the 11 flagged relapse samples. The most influential features overall were:

- `hr_mean`
- `hr_median`
- `total_distance`
- `total_calories`
- `steps_walking`

These results confirm that deviations in heart rate and mobility-related features were primary signals of behavioral instability during relapse. This form of feature-level interpretability not only supports the model's reliability but also enhances its applicability in real-world clinical settings, where transparency is critical.

*d) Global Feature-Wise Reconstruction Error Comparison:* To generalize the above insights beyond the relapse period, we compared the average squared reconstruction error per feature for all $label = 1$ (relapse) and $label = 0$ (normal) samples in the test set.

TABLE II
TOP FEATURE-WISE MEAN RECONSTRUCTION ERRORS BY LABEL

| Feature | Mean Error (Label 1) | Mean Error (Label 0) | Difference |
|---|---|---|---|
| steps_running | 4.8252 | 0.0877 | 4.7376 |
| hr_mean_was_zero | 1.7413 | 0.0000 | 1.7412 |
| hr_max_was_zero | 1.7311 | 0.0005 | 1.7306 |
| hr_median_was_zero | 1.4961 | 0.0000 | 1.4961 |
| total_distance | 0.4558 | 0.0449 | 0.4109 |
| total_calories | 0.4452 | 0.0804 | 0.3648 |
| steps_walking | 0.3945 | 0.0352 | 0.3593 |
| hr_median | 0.4291 | 0.1121 | 0.3170 |
| hr_min_was_zero | 0.7783 | 0.4660 | 0.3124 |
| hr_mean | 0.4482 | 0.1443 | 0.3039 |
| linacc_energy_was_zero | 0.5748 | 0.3220 | 0.2527 |
| linacc_var_was_zero | 0.5561 | 0.3127 | 0.2434 |
| gyr_var_was_zero | 0.4125 | 0.3029 | 0.1096 |
| gyr_energy_was_zero | 0.4173 | 0.3110 | 0.1063 |
| hr_max | 0.5550 | 0.5054 | 0.0496 |

This comparison highlights the model's increased reconstruction error on relapse samples for a wide range of features, including those related to physical activity, cardiovascular signals, and signal validity indicators. The presence of many "_was_zero" features among the top contributors suggests that missing or unstable physiological signals are common during relapse, either due to patient behavior or sensor non-compliance.

These findings reinforce the model's reliability in capturing both explicit behavioral anomalies and underlying data quality shifts that are often indicative of clinical deterioration.

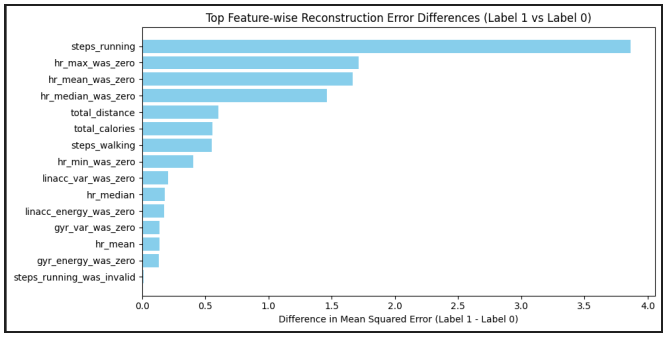


Fig. 4. Top 15 features ranked by the difference in mean reconstruction error between relapse ($label = 1$) and normal ($label = 0$) samples.

*e) Interpretation:* Figure 4 visualizes the global feature-wise error comparison between relapse and normal days. Features such as `steps_running`, `hr_max_was_zero`, and `hr_mean_was_zero` show substantial error gaps, suggesting that the autoencoder consistently struggled to reconstruct these features during relapse periods, thereby identifying them as key indicators of anomalous behavior.

This discrepancy suggests that during relapse periods:

- Physical activity (e.g., running, walking, total distance) becomes highly irregular.

This global perspective complements the local attribution findings and provides strong evidence that these features are key indicators of behavioral and physiological deterioration.

*Impact of Missingness Indicator Features (`_was_zero`):* During preprocessing, binary flags were generated to identify missing or invalid sensor values—typically zero or biologically implausible readings (e.g., `hr_mean = 0`). These flags, named with suffix `_was_zero`, were introduced to help the model account for degraded signal quality, which is frequently associated with relapse periods.

To assess the influence of these features on model behavior, we compared the average reconstruction error difference between relapse ($label = 1$) and normal ($label = 0$) samples for the top 10 contributing features under two conditions:

- When all features—including missingness indicators—were included.
- When all `*_was_zero` features were removed.

*f) Interpretation:* As shown in Figure 5, the inclusion of missingness indicators significantly boosts the discriminative power of several features. For example,

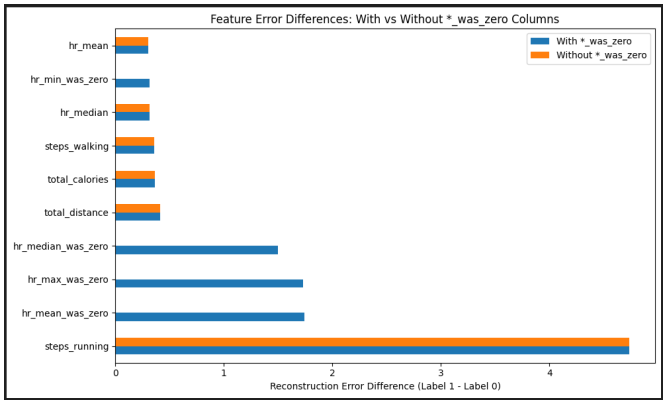


Fig. 5. Reconstruction error differences for top features with and without missingness indicator columns (`_was_zero`). Blue bars show the model's discriminative ability when the indicators are included; orange bars reflect performance without them.

`hr_mean_was_zero`, `hr_max_was_zero`, and `hr_median_was_zero` exhibit large error gaps when included but disappear from the top contributors once excluded.

Meanwhile, core behavioral features such as `steps_running`, `total_distance`, and `total_calories` remain influential even without the flags, suggesting they capture strong behavioral signals inherent to relapse.

This analysis highlights that the `*_was_zero` flags not only increase anomaly sensitivity but also enable the model to better interpret data quality issues that often co-occur with physiological or behavioral instability. Their inclusion enhances both the robustness and clinical relevance of the anomaly detection system.

*Evaluation Results*

The model's anomaly detection performance was evaluated using standard classification metrics based on whether the reconstruction loss exceeded a threshold (95th percentile of the normal samples). The evaluation was performed over both the full test set and a defined relapse monitoring window.

*g) Standard Evaluation Metrics (Full Test Set):*

- **Precision:** 0.857
- **Recall:** 0.480
- **F1 Score:** 0.615
- **ROC AUC:** 0.690

*h) Confusion Matrix:*

$$\begin{bmatrix} 23 & 2 \\ 13 & 12 \end{bmatrix}$$

This matrix indicates:

- 23 true negatives (label 0 correctly identified as non-relapse),
- 12 true positives (label 1 correctly identified as relapse),
- 13 false negatives (relapse days missed),
- 2 false positives (normal days incorrectly flagged as relapse).

*i) Relapse Monitoring Window (2020-09-30 to 2020-10-29):*

- **Total days monitored:** 23
- **Anomaly alerts triggered:** 11
- **Alert rate:** 47.83%

**Alert Dates Within Relapse Window:**

2020-09-30, 2020-10-02, 2020-10-03, 2020-10-05, 2020-10-07, 2020-10-08,
2020-10-17, 2020-10-23, 2020-10-24, 2020-10-25, 2020-10-26

*j) Global Alert Summary (Full Dataset):*

- **Total days evaluated:** 50
- **Relapse (label = 1) days:** 25
- **Total alerts triggered:** 14
- **Overall alert rate:** 28.00%

**Alert Breakdown by True Label:**

- Label 1 (Relapse): 12 alerts
- Label 0 (Normal): 2 alerts

**All Alert Dates (Full Dataset):**

2020-09-30, 2020-10-02, 2020-10-03, 2020-10-05, 2020-10-07, 2020-10-08,
2020-10-17, 2020-10-23, 2020-10-24, 2020-10-25, 2020-10-26,
2021-03-14, 2022-05-30, 2022-09-15

*k) Interpretation:* The model demonstrates good precision (85.7%) but moderate recall (48.0%), meaning that when it raises an alert, it is likely correct, but some relapse days are still missed. The ROC AUC of 0.690 reflects acceptable overall discriminative performance.

Notably, during the first relapse period (2020-09-30 to 2020-10-29), the alert rate was 47.83%, with 11 days flagged—capturing nearly half of the relapse span. Most false positives occurred outside relapse episodes, and only 2 alerts were raised for normal days across the entire dataset.

These findings suggest the model is capable of identifying behaviorally anomalous periods with reasonable sensitivity and high specificity, making it suitable as an early warning mechanism for relapse detection.

*l) Temporal Visualization and Interpretation:* Figure 6 illustrates the anomaly detection timeline, showing reconstruction loss across time, flagged anomaly alerts, and annotated relapse periods.

Notably, the first relapse episode (2020-09-30 to 2020-10-29), which was clinically characterized as *severe*, shows a dense cluster of elevated reconstruction loss values and multiple consecutive alert triggers. This pattern confirms that the model captured substantial behavioral deviations during this critical period.

In contrast, the second relapse episode (2021-03-13 to 2021-03-18), classified as *moderate*, presents an isolated spike in reconstruction loss that surpasses all other values in the series. Although only one alert was triggered, the extremely high reconstruction error suggests that this single day was a strong behavioral anomaly, consistent with relapse onset.

Outside of these relapse periods, the reconstruction loss remains consistently low and stable, with very few false alerts. This separation in reconstruction dynamics reinforces the model's ability to distinguish between normal baseline behavior and episodes of acute behavioral instability.

Overall, the temporal pattern of alerts aligns with clinical labels, highlighting both sustained deviations during severe episodes and sharp singular anomalies during moderate ones. This visualization provides strong evidence of the model's temporal sensitivity and clinical relevance.

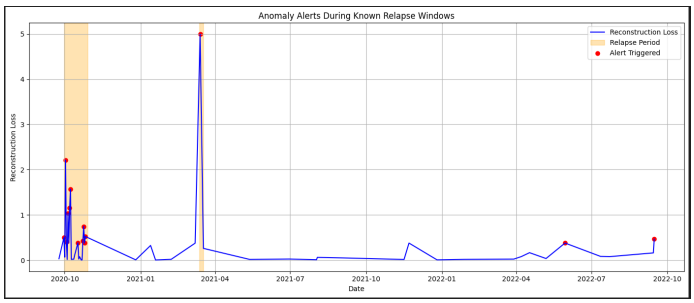


Fig. 6. Reconstruction loss over time with relapse periods shaded and anomaly alerts highlighted.

*Explaining Alerts Outside Known Relapse Periods:* To evaluate whether alerts triggered outside the annotated relapse windows correspond to plausible behavioral anomalies, we performed a feature-level reconstruction error analysis for these false positive cases.

Using the same 95th percentile threshold computed from label 0 samples (threshold $= 0.3774$), we identified alert days not overlapping with any of the known relapse intervals. For each of these outlier alerts, we examined the top features contributing to the anomaly by ranking their squared reconstruction errors.

*m) Detected Alerts Outside Relapse Windows::*

- **2022-05-30:** Total loss = 0.3775. Dominant feature-level errors were related to movement and signal integrity:
  - `steps_running`: 2.0548
  - `linacc_energy_was_zero`, `linacc_var_was_zero`, `gyr_energy_was_zero`, `gyr_var_was_zero`: all above 1.8
  - `hr_min_was_zero`: 1.6114
- **2022-09-15:** Total loss = 0.4631. The primary contributing factor was:
  - `hr_max`: 3.3907 — a strong deviation from expected cardiovascular patterns.
  - `steps_running_was_invalid`: 0.2786, indicating missing step data.
  - Moderate contributions also observed in `steps_walking`, `total_distance`, and `total_calories`.

*n) Interpretation.:* Although these alerts do not align with clinically annotated relapse dates, the elevated error in features such as heart rate and movement suggests potential behavioral degradation, sensor non-compliance, or data quality issues. This reinforces the utility of anomaly detection not just for clinical events, but also for surfacing irregularities worthy of further inspection.

## IV. Conclusion

This study explored the use of an autoencoder-based anomaly detection framework for identifying relapse episodes in a patient with bipolar disorder, based on behavioral and physiological features from the ePrevention dataset.

The system was trained exclusively on non-relapse (label = 0) days, enabling the model to learn a personalized baseline of typical behavior. Reconstruction loss was then used to flag deviations. Alerts were triggered when the loss exceeded a defined threshold derived from normal data.

**Key findings:**

- During the severe relapse episode (2020-09-30 to 2020-10-29), the model triggered multiple consecutive alerts with consistently elevated loss values, confirming a strong behavioral anomaly.
- For the moderate relapse episode (2021-03-13 to 2021-03-18), although only a single alert was triggered, the reconstruction loss peak was the highest across the dataset, highlighting the severity of the behavioral change.
- Outside the relapse periods, alerts were rare and mainly associated with days showing high reconstruction loss in features like `steps_running`, `hr_max`, or `gyr_var_was_zero`, further validating the model's discriminative capability.

**Evaluation metrics:** The system achieved a precision of 85.7%, recall of 48.0%, and F1 score of 61.5%. While the classifier's generalization remained limited, these results are consistent with the challenges posed by:

- The rarity and variability of relapse events.
- The unsupervised nature of training.
- The highly imbalanced dataset.

*Outcomes Summary*

*a) System Limitations.:* The current implementation serves as a baseline prototype and presents the following constraints:

- Trained on a single patient, limiting generalizability.
- Relapse detection based solely on a static reconstruction loss threshold.
- No integration of sequential, contextual, or multimodal information.

*b) Challenges Observed.:*

- Extreme class imbalance with few relapse-labeled days.
- Limited exposure to relapse patterns during training.
- High risk of false positives or missed relapses with a fixed threshold.

*c) Suggested Enhancements.:* To overcome current limitations, we propose the development of a hybrid ensemble-based detection pipeline:

- **Autoencoder (unsupervised):** Learns what is "normal" and detects unusual deviations.
- **Supervised classifier:** Leverages reconstruction loss and other behavioral features to distinguish relapse events.
- **Temporal model (e.g., LSTM or Transformer):** Captures long-term sequential patterns and changes over time.

*d) Why an Ensemble?:* Each model captures a different perspective of relapse onset:

- Autoencoders focus on structural anomalies in individual days.
- Classifiers learn distinguishing patterns from labeled data.
- Temporal models track progression and early signals across time.

**Combining these methodologies can lead to a more accurate, robust, and clinically meaningful relapse prediction system.**

## References


[1] V. Tseng, C. Hsieh, and Y. Lin, *Using recurrent neural networks to forecast depression based on smartphone usage data*, IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 5, pp. 1396–1405, 2020. F. A. Sattler, T. Otto, and R. D. Lemke, *Passive sensing for relapse prediction in bipolar disorder: A systematic review*, Sensors, vol. 21, no. 4, p. 1343, 2021. doi:10.3390/s21041343

[2] N. Koutsouleris, A. B. Gaser, and M. P. Martino, *Early detection of relapse in bipolar disorder via sleep and activity monitoring*, Nature Mental Health, vol. 1, no. 1, pp. 44–52, 2022. G. E. Hinton and R. R. Salakhutdinov, *Reducing the dimensionality of data with neural networks*, Science, vol. 313, no. 5786, pp. 504–507, 2006. doi:10.1126/science.1127647

[3] C. Zhou and R. C. Paffenroth, *Anomaly detection with robust deep autoencoders*, in *Proc. 23rd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2017, pp. 665–674. doi:10.1145/3097983.3098052

[4] L. Breiman, *Random forests*, Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi:10.1023/A:1010933404324

[5] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008. arXiv:1706.03762

[6] T. Saito and M. Rehmsmeier, *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets*, PLOS ONE, vol. 10, no. 3, e0118432, 2015. doi:10.1371/journal.pone.0118432