Towards Temporal Imputation of BDI-II Symptom Trajectories from Irregular Social Media Histories

Julian Rosas Scull
julian.rosas@ciencias.unam.mx
National Autonomous University of Mexico
Mexico City, Mexico

Julio Muñoz-Benítez
jcmunoz@inaoep.mx
Computer Science Department,
National Institute of Astrophysics, Optics and Electronics
Puebla, Mexico

Luis Enrique Sucar
esucar@inaoep.mx
Computer Science Department,
National Institute of Astrophysics, Optics and Electronics
Puebla, Mexico

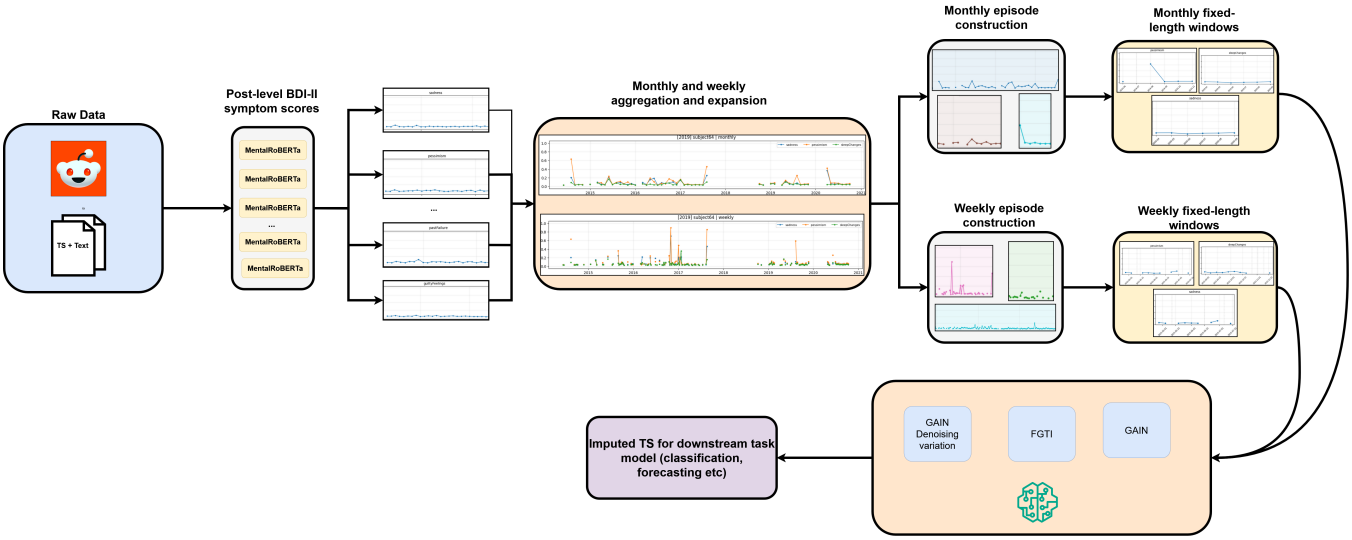

Figure 1: Overview of the proposed framework for temporal imputation of BDI-II symptom trajectories from irregular Reddit histories. The pipeline transforms raw timestamped posts into post-level symptom scores, constructs monthly and weekly temporal windows through aggregation, episode construction, and expansion, and evaluates imputation models to produce complete trajectories for downstream longitudinal analysis.

Abstract

Symptom-level depression detection from social media enables fine-grained analysis of the 21 symptoms defined in the Beck Depression Inventory-II (BDI-II), but post-level predictions do not directly provide complete longitudinal trajectories. Users post irregularly, symptom evidence is sparse, and temporal gaps create substantial missingness. We present a pilot benchmark for temporal imputation of multivariate BDI-II symptom signals derived from Reddit histories. Using the top-10 most active users from the eRisk 2017 and eRisk 2019 datasets, we aggregate MentalRoBERTa-generated post-level symptom scores into monthly and weekly user-level time series, construct temporal episodes, and generate fixed-length windows. To avoid leakage from overlapping windows, we evaluate all methods under a user-level train/test split and an artificial masking protocol over observed test entries. We compare simple statistical baselines, interpolation methods, standard GAIN, a denoising variant of GAIN, and FGTI. Within this small-cohort setting, FGTI obtains the lowest MAE and RMSE in both temporal granularities, while episode-level mean imputation remains a competitive baseline. Denoising GAIN also improves over standard GAIN, suggesting that training-time corruption can provide a useful reconstruction signal in sparse symptom trajectories. Because the imputation targets are classifier-generated symptom scores and the evaluation relies on artificial masking, the results should be interpreted as preliminary evidence for the benchmark design rather than definitive conclusions about clinical symptom dynamics. Overall, this work provides an initial leakage-aware framework for transforming irregular social media-derived symptom predictions into temporally usable trajectories.

CCS Concepts

• **Information systems** → **Data mining**; • **Computing methodologies** → *Machine learning*; *Neural networks*; • **Human-centered computing** → *Social media.*

MiLeTS '26, Jeju, Republic of Korea
2026.

Keywords

Time series imputation, missing data, social media, mental health, BDI-II, eRisk

ACM Reference Format:

Julian Rosas Scull, Julio Muñoz-Benítez, and Luis Enrique Sucar. 2026. Towards Temporal Imputation of BDI-II Symptom Trajectories from Irregular Social Media Histories. In *Proceedings of The 12th Mining and Learning from Time Series Workshop (MiLeTS '26).* ACM, New York, NY, USA, 8 pages.

1 Introduction

Depression is a major public health concern whose clinical presentation involves a broad range of affective, cognitive, behavioral, and somatic symptoms. Standardized instruments such as the Beck Depression Inventory-II (BDI-II) provide a structured way to assess these symptoms [3]. In parallel, computational approaches have increasingly explored social media as a complementary source of information for mental health assessment and monitoring, since platforms such as Reddit contain longitudinal records of how users express emotions, experiences, and mental health concerns over time [5, 7].

Prior work has shown that depression-related language can be modeled not only as a binary user-level label, but also through more interpretable symptom-level representations grounded in the 21 items of the BDI-II [10]. This symptom-level view is important because depressive symptoms are not expressed uniformly in online discourse: some symptoms appear explicitly and frequently, while others are implicit, sensitive, or sparsely reflected in text. Thus, post-level symptom detection provides a useful starting point, but it does not directly yield complete temporal trajectories.

Additionally, missingness is central to this setting. Unlike regularly sampled clinical or sensor measurements, social media observations are available only when users decide to post, so gaps in symptom trajectories may reflect inactivity, lack of textual evidence, platform behavior, or the absence of detectable symptom expression. From the perspective of missing-data mechanisms, such gaps are unlikely to be strictly missing completely at random (MCAR): their occurrence may depend on observed posting behavior and temporal activity patterns, corresponding to missing at random (MAR), and may also depend on unobserved symptom states or willingness to disclose mental-health evidence, corresponding to missing not at random (MNAR)[13]. In this paper, artificial masking is therefore used only as a measurable benchmark protocol over observed entries, not as a claim that real social-media missingness is MCAR.

This paper studies the temporal modeling problem that arises after symptom-level detection. We use Reddit histories from the top-10 most active users in the eRisk 2017 and eRisk 2019 datasets, where each post has been processed with MentalRoBERTa [11] to obtain predicted scores for the 21 BDI-II symptoms [10, 14, 15]. From these post-level predictions, we construct monthly and weekly multivariate symptom trajectories, segment them into temporal episodes, and generate fixed-length windows for imputation. Our goal is not to improve or clinically validate post-level symptom detection, but to evaluate how sparse and irregular classifier-derived symptom signals can be transformed into temporally usable representations.

We propose a pilot benchmark for imputing sparse BDI-II symptom trajectories derived from social media histories. The benchmark includes a temporal preprocessing framework based on aggregation, episode construction, temporal expansion, and fixed-length window generation. To avoid leakage across overlapping windows from the same user, all methods are evaluated under a user-level train/test split. We compare simple statistical baselines, interpolation methods, the standard Generative Adversarial Imputation Nets model (GAIN), a denoising variant of GAIN, and Frequency-aware Generative Models for Multivariate Time Series Imputation (FGTI) [19] under a common artificial masking protocol [21]. Because the retained cohort is small after trajectory filtering, the empirical results are intended to characterize this benchmark setting and to motivate larger-scale validation, rather than to establish a definitive ranking of imputation methods. By evaluating both monthly and weekly temporal granularities, we analyze how temporal construction choices affect the difficulty of imputing social media-derived symptom trajectories.

The remainder of the paper is organized as follows. Section 2 reviews prior work on depression detection from social media and time-series imputation. Section 3 describes the eRisk datasets and the post-level BDI-II symptom scores used in this study. Section 4 presents the temporal preprocessing pipeline used to construct symptom trajectories. Section 5 describes the experimental setup, including the split, masking protocol, compared methods, and evaluation metrics. Section 6 presents and discusses the results. Finally, Section 7 summarizes the main findings and outlines future work.

2 Related Work

Automatic depression detection from social media has been widely studied using user-generated text from platforms such as Reddit, Twitter/X, and online forums [5, 7]. Much of this work formulates the task as binary user-level classification, aiming to infer whether a user is at risk of depression from their posting history. While useful for screening, binary labels provide limited insight into which depressive symptoms are expressed and how they evolve over time. The eRisk benchmark has supported early risk prediction research using Reddit data [14, 15]. More recent work has moved toward symptom-level modeling grounded in standardized instruments such as the Beck Depression Inventory-II (BDI-II), enabling finer-grained evidence of depressive symptom expression [2, 3]. In this direction, Higuera-Gutiérrez et al. evaluated general-purpose transformers, mental-health-adapted models, and LLM-based embeddings for detecting the 21 BDI-II symptoms in social media text [10]. Their findings show that symptoms are unevenly expressed: affective symptoms tend to be more explicit, whereas behavioral or sensitive symptoms are often sparse or underrepresented. Our work builds on this symptom-level perspective, but focuses on the next temporal step: transforming post-level MentalRoBERTa scores into user-level trajectories suitable for longitudinal imputation.

Missing values are a central challenge in time-series analysis, and leaving them untreated can compromise subsequent analysis [12]. Classical approaches such as mean imputation, forward filling, interpolation, and Kalman-based methods remain important baselines because time-series imputation requires exploiting temporal dependencies rather than only inter-attribute correlations [16]. Deep

learning has introduced specialized alternatives: GAIN uses an adversarial generator–discriminator framework [21], SAITS captures temporal dependencies and feature correlations with self-attention [8], and FGTI uses frequency-domain conditions to guide generative imputation in multivariate time series [19]. In our benchmark, we compare statistical and interpolation baselines, standard GAIN, a denoising variant of GAIN, and FGTI, motivated not only by reconstruction accuracy but also by downstream temporal modeling: prior work shows that imputation quality can affect classification and regression on imputed series [1, 4]. In mental-health monitoring, imputation has also been used over irregular and incomplete smartphone-based questionnaire time series to estimate and forecast suicidal ideation, comparing traditional baselines such as LOCF and median imputation with BRITS, Transformer, and SAITS variants [17]. Unlike these settings, our work first constructs sparse BDI-II symptom trajectories from irregular social media histories and evaluates imputation under a leakage-aware user-level split.

3 Dataset

We use Reddit histories from two editions of the CLEF eRisk benchmark. The eRisk 2017 collection corresponds to the Early Detection of Depression task, where systems process chronological sequences of user writings to identify signs of depression as early as possible [14]. The eRisk 2019 collection corresponds to the Measuring the Severity of the Signs of Depression task, where systems estimate depression severity from user posting histories using a questionnaire based on the Beck Depression Inventory [15]. In this work, we initially select the top-10 most active users from each dataset, yielding 20 candidate users for a small but dense pilot subset. This choice is motivated by the temporal nature of our study, since users with denser posting histories are more likely to provide enough observations for constructing monthly and weekly symptom trajectories.

Each post is associated with a timestamp and a 21-dimensional vector of BDI-II symptom scores produced by symptom-specific MentalRoBERTa classifiers, following prior work that formulates depressive symptom detection as 21 independent BDI-II classification problems [10, 11]. These values should be interpreted as model-derived symptom scores rather than clinically observed symptom labels. Errors or biases in the upstream symptom classifiers can therefore propagate into the constructed trajectories and into the imputation targets. Because users post at irregular intervals and long periods may contain no textual evidence, these post-level scores do not directly provide complete trajectories. We therefore construct monthly and weekly multivariate time series from these histories, segment them into activity episodes, and extract fixed-length windows for imputation. Candidate users are retained only if at least one constructed episode is long enough to generate a valid window under the selected monthly or weekly settings; after this filtering step, the final benchmark contains the users reported in Table 2. All experiments are evaluated under a user-level train/test split, with disjoint users assigned to each split.

4 Methodology: From Posts to Temporal Symptom Trajectories

We describe a preprocessing framework for transforming irregular social media histories into temporal symptom trajectories for imputation. The input consists of timestamped Reddit posts associated with 21 BDI-II symptom scores, while the target representation is a set of fixed-length multivariate windows that preserve each user's longitudinal structure. The main challenge is that users post at irregular intervals, so post-level predictions do not directly form complete time series. Our framework addresses this through four steps: aggregating post-level scores into monthly and weekly views, segmenting user histories into activity episodes, expanding each episode into a regular timeline with explicit missing values, and generating fixed-length windows for imputation.

4.1 Problem formulation

Let u denote a user and let N_u be the number of posts authored by that user. For each post i, we observe a timestamp t_i and a vector of predicted BDI-II symptom scores $\mathbf{s}_i \in \mathbb{R}^{21}$, where each component corresponds to one depressive symptom. The post history of user u is represented as the ordered sequence

$$P_u = \big((t_1, \mathbf{s}_1), \ldots, (t_{N_u}, \mathbf{s}_{N_u})\big), \quad t_1 \le \cdots \le t_{N_u}. \tag{1}$$

Since posts are irregularly spaced, P_u does not directly define a regularly sampled time series. We therefore transform each user history into a multivariate temporal trajectory $X_u^{(g)} \in \mathbb{R}^{T_u \times 21}$, where $g \in \{\text{monthly}, \text{weekly}\}$ is the temporal granularity and T_u is the number of time buckets in the constructed timeline. A bucket with posts is assigned aggregated symptom scores, while a bucket without posts is treated as missing rather than zero, because the absence of posts indicates lack of textual evidence rather than absence of depressive symptoms.

We define a binary observation mask $M_u^{(g)} \in \{0,1\}^{T_u \times 21}$, where $M_{u,t,j}^{(g)} = 1$ if symptom j has an observed aggregated value for user u at time bucket t, and 0 otherwise. The imputation task is to estimate the missing components of $X_u^{(g)}$ from the observed values and the temporal context of the user trajectory. This component-wise formulation is consistent with standard imputation settings, where models estimate missing entries from the observed components of the data vector or sequence [21].

4.2 Temporal Aggregation

We convert each user's irregular posting history into two regular temporal views: one organized by calendar month and another by calendar week. These views are treated as separate datasets, not as direct conversions of one another. The monthly view provides a coarser but denser trajectory, while the weekly view preserves finer temporal variation at the cost of introducing more missing periods.

For a given granularity g, let $P_{u,t}^{(g)}$ denote the set of posts written by user u within time interval t. If this set is non-empty, the value of symptom j at that interval is computed as the average post-level score:

$$X_{u,t,j}^{(g)} = \frac{1}{|P_{u,t}^{(g)}|} \sum_{i \in P_{u,t}^{(g)}} s_{i,j}. \tag{2}$$

If $P_{u,t}^{(g)}$ is empty, then $X_{u,t,j}^{(g)}$ is left missing for all symptoms j. Thus, a period with posts becomes an observed 21-dimensional symptom vector, while a period without posts becomes a missing time step. This preserves the distinction between low predicted symptom scores and absence of textual evidence.

4.3 Episode Construction

After aggregation, each user trajectory may contain long inactive periods with no observed time buckets. We therefore split trajectories into episodes, defined as contiguous periods of user activity separated by extended gaps. Given two consecutive observed buckets b_k and b_{k+1}, a new episode is started when their gap exceeds an inactivity threshold τ_g. The threshold is defined in the same unit as the temporal granularity: months for monthly trajectories and weeks for weekly trajectories. This prevents imputation from connecting symptom observations that are separated by long periods without textual evidence.

4.4 Temporal Expansion and Window Generation

After episodes are identified, each episode is expanded into a complete regular timeline. For an episode e, we include every monthly or weekly bucket between its first and last observed bucket. Buckets with posts retain their aggregated 21-dimensional symptom vector, while buckets without posts are inserted as missing rows. This produces an episode-level matrix $X_e^{(g)} \in \mathbb{R}^{T_e \times 21}$, together with a binary observation mask $M_e^{(g)} \in \{0,1\}^{T_e \times 21}$, where observed entries are marked with 1 and missing entries with 0.

The expanded episodes are then divided into fixed-length sliding windows. Each window contains L_g consecutive time buckets and all 21 symptom dimensions. For example, if a monthly episode contains eight months and $L_g = 6$, the resulting windows cover months 1–6, 2–7, and 3–8. Thus, each monthly window contains six consecutive monthly symptom vectors, and each weekly window analogously contains L_g consecutive weekly symptom vectors.

Formally, a window starting at position r is defined as

$$W_{e,r}^{(g)} = X_e^{(g)}[r : r + L_g - 1, :]. \tag{3}$$

where r indexes the starting bucket, L_g is the window length for granularity g, and : indicates that all 21 symptom dimensions are retained. Each window therefore has shape $L_g \times 21$. The corresponding mask window is extracted in the same way:

$$A_{e,r}^{(g)} = M_e^{(g)}[r : r + L_g - 1, :]. \tag{4}$$

Episodes shorter than L_g time buckets cannot produce a window and are therefore excluded from window generation. All valid sliding windows produced from the remaining episodes are retained, together with their corresponding observation masks.

5 Experimental Setup

This section defines the benchmark used to evaluate imputation over the constructed symptom trajectories. We describe the selected temporal configurations, the user-level train/test split, the artificial masking protocol, the compared methods, and the evaluation metrics. These choices allow us to compare imputation approaches under a common setting while avoiding leakage across users and preserving the sparse nature of the social media-derived trajectories.

5.1 Trajectory Construction Settings

The temporal trajectories depend on two preprocessing choices: the inactivity threshold used to split episodes and the window length used to produce fixed-size samples. We selected these settings separately for the monthly and weekly views after a diagnostic sweep over candidate thresholds and window lengths. The sweep considered the number of eligible users, generated windows, and observed-value fraction, aiming to retain enough temporal context without producing overly sparse samples.

Setting	Monthly	Weekly
Inactivity threshold	4 months	4 weeks
Window length	6 months	12 weeks
Window shape	6×21	12×21

Table 1: Selected trajectory construction settings.

Table 1 summarizes the final settings. The weekly view is treated as a separate higher-resolution representation of the same user histories, not as a direct calendar conversion of the monthly view.

Gran.	Split	Users	Windows	Obs. frac.
Month	Train	12	352	0.882
Month	Test	3	77	0.840
Week	Train	12	929	0.864
Week	Test	3	188	0.873

Table 2: Final temporal samples under the selected settings.

Table 2 reports the resulting samples after aggregation, episode construction, temporal expansion, and window generation. The same user-level train/test split is used for both granularities, so the difference in sample counts is due to the temporal construction rather than different user assignments. As expected, the weekly view yields more windows because it represents each history at a finer temporal resolution.

5.2 User-level Train/Test Split

Because multiple overlapping windows can be generated from the same user, we split the data by user rather than by window. This choice is consistent with the user-level formulation commonly used in social-media mental-health modeling, where labels and predictions are associated with users and their posting histories rather than with independent posts [6, 20]. All windows from a user are assigned exclusively to either training or test, preventing evaluation on windows derived from users already seen during training. The same split is used for monthly and weekly trajectories. After excluding users without valid windows, the final split contains 12 training users and 3 test users, producing the sample counts in Table 2.

5.3 Artificial Masking Protocol

To evaluate imputation performance, we artificially mask observed entries in the test windows. Naturally missing entries do not have ground-truth values, so they cannot be used to compute reconstruction error. We therefore randomly hide 20% of the entries that are originally observed in each test set and provide each method with the partially masked window and its updated observation mask. Imputation error is computed only on these artificially hidden entries, following the common evaluation strategy used in imputation benchmarks where known values are removed to create measurable targets [19, 21]. This protocol should be interpreted as an evaluation proxy rather than a complete simulation of real social-media missingness. In particular, random point masking does not capture all forms of behavior-dependent or contiguous gaps that may occur when users stop posting or stop disclosing symptom-relevant content.

Let X_n and M_n denote the n-th test window and its original observation mask, where $M_{n,t,j} = 1$ indicates that symptom j is observed at time step t. We define an evaluation mask H_n, where $H_{n,t,j} = 1$ if an originally observed entry is selected to be hidden for evaluation, and 0 otherwise. The mask given to the imputation model is

$$\widetilde{M}_n = M_n \odot (1 - H_n). \tag{5}$$

so entries selected by H_n are treated as missing at inference time. The original values of these entries are retained only for evaluation, ensuring that each imputed value is compared against a known symptom score.

5.4 Comparison of Methods

We compare imputation methods ranging from simple baselines to deep generative models. The baselines provide reference points for assessing whether learning-based methods improve over basic statistical or temporal assumptions. We include forward-fill, which propagates the most recent observed value within each window and, when needed, uses the nearest future value for leading missing entries. We also include linear interpolation, which estimates missing values from neighboring observations along the temporal axis. In addition, we evaluate user mean, which fills missing entries using symptom-level averages estimated from the corresponding user. Finally, we include episode mean, which fills missing entries using the average symptom value within the corresponding activity episode. This baseline captures local episode-level tendencies and provides a useful reference for assessing whether learned models improve over stable within-episode averages.

We also evaluate GAIN, a generative adversarial imputation model in which a generator completes missing components and a discriminator predicts which components were observed or imputed [21]. In addition to the standard GAIN formulation, we evaluate a denoising variant designed for sparse temporal symptom trajectories. During training, this variant randomly hides a subset of originally observed entries and treats them as reconstruction targets. The generator is therefore trained not only to complete naturally missing entries, but also to recover observed values that were artificially corrupted during training. This follows the denoising principle of learning to reconstruct clean targets from corrupted inputs [18], a strategy also explored for missing-data imputation with denoising autoencoders [9], while retaining the adversarial imputation structure of GAIN [21].

Finally, we include FGTI, a frequency-aware generative model for multivariate time-series imputation. FGTI introduces frequency-domain conditions, including high-frequency and dominant-frequency components, to guide generative imputation and improve the reconstruction of temporal patterns [19]. All methods are evaluated under the same train/test split, trajectory construction settings, and artificial masking protocol.

5.5 Evaluation Metrics

We evaluate imputation accuracy using mean absolute error (MAE) and root mean squared error (RMSE), computed only over the artificially masked test entries. MAE measures the average absolute deviation between imputed and ground-truth symptom scores, while RMSE gives larger weight to larger errors. Lower values indicate better performance.

Let Ω_{test} be the set of artificially masked test entries. For ground-truth values $X_{n,t,j}$ and imputed values $\widehat{X}_{n,t,j}$, we compute

$$\text{MAE} = \frac{1}{|\Omega_{\text{test}}|} \sum_{(n,t,j)\in\Omega_{\text{test}}} \left| X_{n,t,j} - \widehat{X}_{n,t,j} \right|. \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{1}{|\Omega_{\text{test}}|} \sum_{(n,t,j)\in\Omega_{\text{test}}} \left(X_{n,t,j} - \widehat{X}_{n,t,j} \right)^2}. \tag{7}$$

5.6 Implementation Details

All experiments were conducted on the Kaggle platform using two NVIDIA T4 GPUs. We used the same user-level train/test split, trajectory construction settings, and artificial masking protocol for all methods. For the learning-based models, we used the default hyperparameters from the original implementations whenever possible, modifying only those required to match our data representation. In particular, the input and output dimensions were adapted to the 21 BDI-II symptom variables, the sequence length was set to the selected monthly or weekly window length, and FGTI frequency-related settings were adjusted to the shorter windows used in our benchmark. For GAIN and Denoising GAIN, the same architecture and training configuration were used, with the denoising variant adding only the training-time random masking described above.

6 Results and Discussion

Table 3 reports the test imputation results for the monthly and weekly settings. Within this pilot split and masking draw, FGTI obtains the lowest MAE and RMSE in both granularities. In the monthly setting, FGTI is followed by Denoising GAIN in MAE and by the episode-level mean baseline in RMSE. In the weekly setting, FGTI also achieves the lowest error, with a clearer separation from the remaining methods. The monthly and weekly evaluations contain 1,605 and 8,211 artificially masked test entries, respectively. Because these entries come from overlapping windows generated from only three held-out users, the results should be interpreted as descriptive evidence for this benchmark configuration rather than as statistically definitive model rankings.

The advantage of FGTI is more pronounced in MAE than in RMSE. In terms of MAE, FGTI improves over the second-best

Gran.	Method	MAE	RMSE
Month	FGTI	**0.0116**	**0.0338**
Month	Denoising GAIN	0.0172	0.0380
Month	Episode mean	0.0191	0.0364
Month	Linear interp.	0.0197	0.0407
Month	Forward-fill	0.0201	0.0409
Month	User mean	0.0212	0.0425
Month	GAIN	0.0334	0.0518
Week	FGTI	**0.0101**	**0.0355**
Week	Episode mean	0.0192	0.0425
Week	Denoising GAIN	0.0193	0.0439
Week	Linear interp.	0.0207	0.0499
Week	User mean	0.0212	0.0469
Week	Forward-fill	0.0220	0.0540
Week	GAIN	0.0296	0.0505

Table 3: Test imputation performance under monthly and weekly temporal granularities. Lower is better.

Figure 2: Qualitative comparison under the weekly temporal setting for subject2714, *pessimism*. Black points denote observed values, green points denote artificially masked ground-truth values, red points denote model predictions at masked locations, and the colored lines denote the completed imputed trajectories.

method by approximately 33% in the monthly setting and 47% in the weekly setting. These relative gains are useful for describing the observed pattern, but they should not be read as significance-tested improvements because the current benchmark has a small number of held-out users and does not evaluate multiple train/test splits or masking draws. Standard GAIN performs poorly relative to both simple baselines and Denoising GAIN, especially in the monthly setting. This suggests that the original adversarial imputation objective may be difficult to optimize in this small-data regime, where the number of users is limited and many windows are derived from overlapping segments of the same trajectories. In contrast, the denoising variant improves over standard GAIN in both granularities, reducing MAE by about 49% in the monthly setting and 35% in the weekly setting. This pattern is consistent with the idea that training-time corruption of observed entries can provide an auxiliary reconstruction signal, although larger-scale experiments are needed to assess its robustness.

The baseline results provide useful context for interpreting the learned models. Forward-fill and linear interpolation rely on local temporal continuity within each window, user mean provides a global symptom-level reference estimated from the training set, and episode mean provides an episode-level reference computed from the corresponding evaluated episode. The competitive performance of these baselines suggests that the constructed trajectories contain local smoothness and episode-level stability. The lower error obtained by FGTI within this setting suggests that it may capture temporal or cross-symptom structure beyond average-level trends, but this interpretation remains provisional given the small cohort and the use of classifier-generated targets.

6.1 Effect of Temporal Granularity

The selected trajectory settings produce two complementary views of the same user histories. The monthly setting generates fewer windows, but each time step summarizes a longer period of activity. In contrast, the weekly setting yields more windows and evaluation targets, while representing symptom variation at a finer temporal resolution. Thus, the weekly benchmark provides a larger evaluation set, but also exposes the methods to a more irregular and fine-grained temporal representation.

The relative ordering of methods is mostly stable across granularities in this pilot evaluation: FGTI achieves the lowest error in both settings, Denoising GAIN improves over standard GAIN, and simple temporal baselines remain competitive. However, RMSE increases for several methods in the weekly setting, suggesting that finer temporal resolution can introduce harder reconstruction cases. This is expected in social media trajectories, where a user may have enough posts to support a monthly symptom estimate but still show sparse or uneven evidence at the weekly level.

Figures 2 and 3 provide qualitative examples for the same user and symptom under the weekly and monthly settings, respectively. Both figures show the *pessimism* trajectory for subject2714 from the 2017 dataset under FGTI, Denoising GAIN, and GAIN. These examples are intended to illustrate typical reconstruction behavior rather than to serve as independent evidence of model superiority: FGTI and Denoising GAIN place predictions close to several masked ground-truth values, while standard GAIN appears less stable in regions with sharper local changes. The weekly example illustrates the finer and more irregular temporal representation, whereas the monthly example shows a more compact trajectory with fewer time steps.

6.2 Scope and Limitations

This study should be interpreted as a pilot benchmark rather than a definitive comparison of imputation methods for mental-health trajectories. The final retained cohort contains 15 users, with 12

Figure 3: Qualitative comparison under the monthly temporal setting for subject2714, *pessimism*. Black points denote observed values, green points denote artificially masked ground-truth values, red points denote model predictions at masked locations, and the colored lines denote the completed imputed trajectories.

users for training and 3 users for testing after temporal filtering. Although the split is leakage-aware at the user level, the test windows are still overlapping samples derived from a small number of trajectories. As a result, the reported errors and relative improvements should be understood as descriptive results for this setting, without claims of statistical significance or population-level generalization.

A second limitation is that the imputation targets are MentalRoBERTa-generated BDI-II symptom scores rather than clinically validated symptom measurements. The benchmark therefore evaluates the reconstruction of model-derived symptom signals, and any upstream classifier error may affect both the trajectories and the measured imputation error. Finally, the artificial masking protocol hides random observed entries to create measurable targets. This is standard in imputation evaluation, but it does not fully reproduce the behavior-dependent, potentially contiguous, and possibly MNAR missingness mechanisms expected in social media. Larger cohorts, repeated masking draws, alternative user splits, and block-based masking protocols are needed to test the robustness of the observed trends.

7 Conclusions and Future Work

We presented a pilot benchmark for imputing sparse BDI-II symptom trajectories derived from Reddit histories in the eRisk 2017 and eRisk 2019 datasets. Starting from post-level MentalRoBERTa symptom scores, we constructed monthly and weekly multivariate trajectories through aggregation, episode construction, temporal expansion, and fixed-length window generation. To reduce leakage from overlapping windows, all methods were evaluated under a user-level train/test split and a common artificial masking protocol.

Within this pilot setting, FGTI obtains the lowest MAE and RMSE in both monthly and weekly evaluations, while episode-level mean imputation remains a competitive baseline. Denoising GAIN also improves over standard GAIN, suggesting that training-time corruption of observed entries may provide a useful reconstruction signal in sparse symptom trajectories. These findings should be interpreted as preliminary because the retained cohort is small, the test set contains only three users, and the imputation targets are classifier-derived symptom scores rather than clinically validated symptom measurements.

Overall, this paper provides an initial step toward transforming irregular symptom-level social media predictions into temporally usable representations for future longitudinal mental-health modeling. Future work should evaluate the proposed pipeline on larger user cohorts, additional eRisk editions, repeated masking draws, alternative user-level splits, and leave-one-user-out settings. It should also consider block or burst masking protocols that better reflect contiguous periods of missing evidence in social media histories, and should examine whether imputed symptom trajectories improve downstream tasks such as user clustering, trajectory characterization, or future risk prediction.

References

[1] Karan Aggarwal and Jaideep Srivastava. 2023. Filling out the Missing Gaps: Time Series Imputation with Semi-Supervised Learning. *arXiv preprint arXiv:2304.04275* (2023).
[2] Mario Ezra Aragón, A. Pastor López-Monroy, Manuel Montes-y Gómez, and David E. Losada. 2025. Online Expressions, Offline Struggles: Using Social Media to Identify Depression-Related Symptoms. *Online Social Networks and Media* 50 (2025), 100338.
[3] Aaron T. Beck, Robert A. Steer, and Gregory K. Brown. 1996. *Manual for the Beck Depression Inventory-II.* Psychological Corporation, San Antonio, TX.
[4] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. 2018. BRITS: Bidirectional Recurrent Imputation for Time Series. In *Advances in Neural Information Processing Systems*, Vol. 31.
[5] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review. *npj Digital Medicine* 3, 1 (2020), 43.
[6] Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1485–1497.
[7] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media.*
[8] Wenjie Du, David Côté, and Yan Liu. 2023. SAITS: Self-Attention-based Imputation for Time Series. *Expert Systems with Applications* 219 (2023), 119619.
[9] Lovedeep Gondara and Ke Wang. 2017. MIDA: Multiple Imputation using Denoising Autoencoders. *arXiv preprint arXiv:1705.02737* (2017).
[10] Cielo Aholiva Higuera-Gutiérrez, Irvin Hussein López-Nava, Manuel Montes-y Gómez, Mario Ezra Aragón, and David E. Losada. 2026. Automatic Depressive Symptom Detection on Social Media using the BDI-II. In *MCPR.* Accepted paper.
[11] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. *arXiv preprint arXiv:2110.15621* (2021).
[12] Mourad Khayati, Alberto Lerner, Zakhar Tymchenko, and Philippe Cudré-Mauroux. 2020. Mind the Gap: An Experimental Evaluation of Imputation of Missing Values Techniques in Time Series. *Proceedings of the VLDB Endowment* 13, 5 (2020), 768–782. doi:10.14778/3377369.3377383
[13] Roderick J. A. Little and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data* (3 ed.). Wiley.
[14] David E. Losada, Fabio Crestani, and Javier Parapar. 2017. eRisk 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 346–360.
[15] David E. Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of eRisk at CLEF 2019: Early Risk Prediction on the Internet. In *CLEF Working Notes.*
[16] Steffen Moritz, Alexis Sardá, Thomas Bartz-Beielstein, Martin Zaefferer, and Jörg Stork. 2015. Comparison of Different Methods for Univariate Time Series

Imputation in R. *arXiv preprint arXiv:1510.03924* (2015).

[17] Gwenolé Quellec, Sofian Berrouiguet, Margot Morgiève, Jonathan Dubois, Marion Leboyer, Guillaume Vaiva, Jérôme Azé, and Philippe Courtet. 2024. Predicting suicidal ideation from irregular and incomplete time series of questionnaires in a smartphone-based suicide prevention platform: a pilot study. *Scientific Reports* 14, 1 (2024), 20870. doi:10.1038/s41598-024-71760-1

[18] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*. ACM, 1096–1103.

[19] Xinyu Yang, Yu Sun, Xiaojie Yuan, and Xinyang Chen. 2024. Frequency-aware Generative Models for Multivariate Time Series Imputation. In *Advances in Neural Information Processing Systems*, Vol. 37.

[20] Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and Self-Harm Risk Assessment in Online Forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2968–2978.

[21] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GAIN: Missing Data Imputation using Generative Adversarial Nets. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. 5689–5698.