Improving Forecasts of Suicide Attempts for Patients with Little Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Ecological Momentary Assessment provides real-time data on suicidal thoughts and behaviors, but predicting suicide attempts remains challenging due to their rarity and patient heterogeneity. We show that single models fit to all patients perform poorly, while individualized models overfit with limited data. To address this, we introduce a Latent Similarity Gaussian Process (LSGP) that models patient heterogeneity, enabling those with little data to leverage similar patients' trends. Preliminary results show improved sensitivity over baselines and offer new understanding of patient similarity.

1 Introduction and Related Work

Ecological Momentary Assessment (EMA) studies leverage smartphones and wearable sensors to capture insights into suicidal thoughts and behaviors (STBs) as they unfold in daily life [1]. In these intensive longitudinal studies, patients are surveyed multiple times daily on their suicidal urges, intent, and affects. This presents opportunities for machine learning (ML) to forecast imminent suicide risk in time for intervention; however, to date, no current approach can do this reliably [2].

Prior work primarily focuses on forecasting suicidal ideation from EMA data (e.g. [3–6]). While forecasting ideation is itself challenging, suicide attempts are even harder to predict due to their low base-rate [7]; even in largest datasets (e.g. 600 patients), attempts are rarely captured (e.g. [8]). This severely limits the data available for model training and evaluation. To exacerbate this challenge, recent work shows that patients' paths to suicide ideation are heterogeneous, suggesting that, at the very least, there are many subtypes of at-risk patients, advocating against the use of single models across all patients [9–11], further reducing the number of data points per model.

In this work, we show that the same patient heterogeneity found in suicidal ideation is found in suicidal attempts. We then present a single model to improve forecasts for patients with little data by capturing patient heterogeneity. Our contributions are:

(A) As with suicidal ideation, we show that a single model trained on data to predict suicide attempts from all patients performs worse than individualized, per-patient models. Specifically, we show that each patient exhibits a different forecasting trend, that, when combined, conflict with one another, resulting in poor forecasting performance. This underscores the importance of explicitly modeling patient heterogeneity [11, 10]. From these results, we may be tempted to use a different model per patient—but per-patient models are prone to severe overfitting for patients with little data.

(B) We naturally formalize our observations into a single model to capture patient heterogeneity, grounded in modeling assumptions supported by our analysis and prior work. Our Latent Similarity Gaussian Process (LSGP) posits that patients lie in a latent space in which distance corresponds to similarity in forecasting trends. By inferring patients' locations in this latent space,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

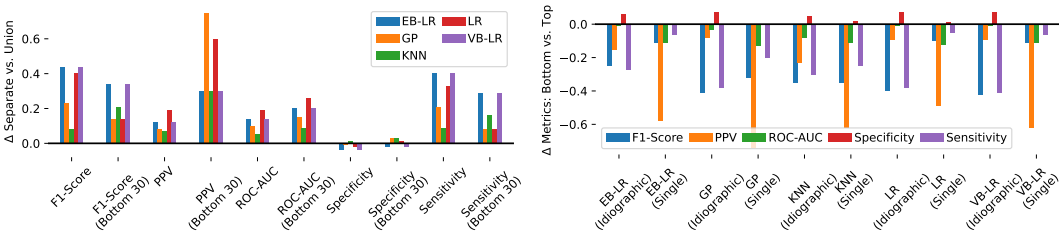


Figure 1: **Left: Idiographic models outperform their counterparts**—except for specificity, which stayed constant, the magnitude of difference in metrics between the idiographic and single models is always positive. **Right: The 30% patients with fewest recorded SREs consistently receive worse forecasts across nearly all metrics and all models**—the magnitude of difference in metrics computed for all patients vs. for the bottom 30% is always negative, indicating lower performance.

forecasts for patients with little data intelligently draw on trends from similar patients. While inspired by prior methods (see Section 3), LSGPs have never been previously applied in this context.

(C) Our preliminary results show promise in improving forecasts of suicide attempts from EMA data for patients with little data, and reveal new avenues for understanding patient similarity. Our approach matches baseline performance on most metrics and notably outperforms baselines in sensitivity, which is especially critical for suicide prevention. Furthermore, we introduce a graph-based visualization of patient similarity within the learned latent space, offering interpretable insights into individualized risk profiles and potential shared mechanisms.

2 The Geometry of Forecasts for At-Risk Patients

Notation. Let N denote the number of patients in the data. Let $x_i = [n_i \quad r_{i,1} \quad \dots \quad r_{i,D_x}]^\intercal$ represent the ith observation in the data, belonging to patient n_i at time t_i, consisting of their responses $r_{1,d} \in \{0, \dots, 10\}$ to 10-point likert-scale EMA questions. Here, we will use questions about patients' affects, suicidal intent/urge and behaviors—for details on the study and data, see Appendix A. Using patient responses to these questions, our task is to predict $y_i \in \{0, 1\}$—whether patient n_i engaged in any suicide related event (SRE) sometime in the week following t_i. We define an SRE as either a self-injurious behavior with some (non-zero) intention of dying, or a presentation to a hospital with suicidal thoughts to prevent the occurrence of a suicide attempt. Let $\mathcal{D} = X, Y$ represent the entire training data. Let $\mathcal{D}_n = X_n, Y_n$ represent patient n's training data, where $X_n = \{x_i | n_i = n\}$ and $Y_n = \{y_i | n_i = n\}$. Note that every patient has a different amount of data.

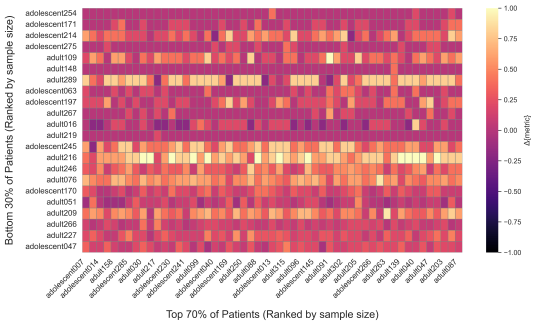
Goal. Given $\mathcal{D}$, our goal is to predict whether, given a *new* EMA response, x_n^*, patient n will engage in an SRE sometime in the next week, y_n^*.

Single vs. Idiographic Models. To better understand the geometry of patient classification boundaries, we compare models trained on *all* patient data $(y_n^* | x_n^*, \mathcal{D})$ with a model consisting of a collection of models—*one per patient* $(y_n^* | x_n^*, \mathcal{D}_n)$. We refer to the former and latter as a single and an idiographic model, respectively. If idiographic models consistently outperform the single models, this suggests that patients have differing forecasting trends. Even before comparing their performance, we note that idiographic models have one major shortcoming: they cannot be used to make predictions for a new patient n^*; we address this limitation in our method (Section 3).

Baselines and Metrics. We compare our method with several baselines, each used both as a single and idiographic model: Gaussian Process Classification (GP) with a Laplace Approximation, k-Nearest Neighbor Classifier (KNN), Logistic Regression (LR), and Bayesian LR with an empirical Bayes type II and variational (EB-LR and VB-LR) approximations. For evaluation, we use: F1-Score, Positive Predictive Value (PPV), Area Under the Receiver Operating Characteristic Curve (ROC-AUC), Specificity, and Sensitivity.

Finding: Patients exhibit conflicting classification boundaries. Fig. 1 (left) shows that, across all baselines and metrics (except for specificity), using an idiographic model results in significantly better performance than the single model. Based on these results, we may be tempted to just use separate models; however, as Fig. 1 (right) shows, both single and idiographic models make worse forecasts for the patients with fewest recorded SREs.

Single model predictions are most influenced by patients with more data, generalizing poorly to patients with less data; idiographic model overfit to patients with little data. We may be further tempted to find *group* patients together that share similar trends, but this is challenging when data are sparse. The figure on the right shows how supplementing the training data of patients with few recorded suicide attempts—by adding data from others—can improve, worsen, or have no effect on their forecasting sensitivity. These mixed results highlight the difficulty of creating patient groups by naively matching patients with fewer observations with those with more. To address this, we next propose a model that leverages the full dataset to capture patient similarity more effectively.



3 Method

There are many pathways to suicide; even among mental health disorders, conditions such as major depression, generalized anxiety, post-traumatic stress, and borderline personality disorder each present distinct mechanisms leading to elevated suicide risk [12, 13]. Moreover, within a single diagnosis, each patient's unique life circumstances—e.g. shaped by social determinants and individual differences—further contributes to patient heterogeneity [10]. To capture this probabilistically, we must allow each patient to follow an individual forecasting trajectory while enabling those with limited data to leverage information from others without imposing a one-size-fits-all solution. We address this by embedding patients in a latent space, where proximity reflects similarity in risk trajectories; forecasts for patients with little data can thus intelligently borrow strength from their nearest neighbors.

Latent Similarity Gaussian Processes (LSGPs). We naturally arrive at the model,

$$\begin{aligned} &\text{Ⓐ}\ z_n \sim p(z) = \mathcal{N}(0, \mathbb{I}_{D_z}), \qquad &&\text{Ⓑ}\ \hat{x}_i | x_i, Z = [r_{i,1} \quad \dots \quad r_{i,D_x} \quad z_{n_i,1} \quad \dots \quad z_{n_i,D_z}]^\intercal, \\ &\text{Ⓒ}\ F|\widehat{X}; \theta \sim \mathcal{N}(0, K_\theta(\widehat{X}, \widehat{X})), \qquad &&\text{Ⓓ}\ y_i | f_i \sim \text{Bernoulli}(\text{sigmoid}(f_i)), \end{aligned}$$

where $\hat{x}_i$ represents the concatenation of the inputs x_i with the latent variable z_{n_i} corresponding to patient n_i, $\widehat{X}$ is a matrix consisting of all x_i's as rows, $K_\theta(\cdot, \cdot)$ is the kernel matrix computed on rows of its arguments with hyperparameters θ, F is a concatenation of all function values f_i corresponding to each x_i, and $\mathbb{I}_{D_z}$ is an identity matrix of width D_z.

Related Models. Our model bears similarity to several existing models, including (i) GP with Latent Covariate [14] or Covariate GP Latent Variable Models [15], but adapted to have *multiple* observations per latent variable, (ii) a Multi-Group GPs [16], but in which the "group" is both *continuous and latent*, or (iii) a Meta-Learning GPs [17], but without the *control signal*.

Sparse Variational LSGPs. Analytical inference is impossible due to the non-Gaussianity of the likelihood and the large number of observations (14763 from $N = 77$ patients), so we apply the sparse variational formulation of GPs [18] to our model, replacing Ⓒ above with:

$$\begin{aligned} &\text{Ⓒ.1}\ U; W, \theta \sim \mathcal{N}(0, K_\theta(W, W)), \\ &\text{Ⓒ.2}\ F|U, X, Z; W, \theta \sim \mathcal{N}(\Psi \cdot U, K_\theta(\widehat{X}, \widehat{X}) - \Psi \cdot K_\theta(\widehat{X}, W)^\intercal), \end{aligned}$$

where $\Psi = K_\theta(\widehat{X}, W) \cdot K_\theta(W, W)^{-1}$. In this formulation, $W \in \mathbb{R}^{M \times (D_x + D_z)}$ is a matrix of M inducing point locations used to "summarize" the training data, enabling more efficient inference.

Stochastic Variational Inference (SVI). We learn W, θ by minimizing the divergence between an approximate and true posterior [18]:

$$W^*, \theta^*, \phi^* = \text{argmin}_{W,\theta,\phi} D_{\text{KL}} \left[q(F, U, Z; W, \theta, \phi) || p(F, U, Z | U, X, Z, Y; W, \theta) \right], \tag{1}$$

using the variational family, $q(F, U, Z; W, \theta, \phi) = p(F|U, X, Z; W, \theta) \cdot q(U; \phi) \cdot \prod_{n=1}^{N} q(z_n; \phi)$, where ϕ are the parameters of full-covariance Gaussian $q(U; \phi) = \mathcal{N}(\mu_\phi, \Sigma_\phi)$ and mean-field

Table 1: **Comparison of Methods on Test Metrics.** We report sensitivity by stratifying patients into bottom, middle, and top thirds based on SRE count.

		ROC-AUC	PPV	Specificity	Sensitivity *Bottom 33%*	Sensitivity *Middle 33%*	Sensitivity *Top 33%*
Single	**RBF-GP**	0.74 ± 0.00	0.66 ± 0.04	0.98 ± 0.00	0.01 ± 0.02	0.36 ± 0.01	0.21 ± 0.01
	KNN	0.70 ± 0.01	0.61 ± 0.03	0.97 ± 0.00	0.03 ± 0.04	0.34 ± 0.01	0.28 ± 0.03
	LR	0.68 ± 0.01	0.60 ± 0.06	$\mathbf{0.99 \pm 0.00}$	0.02 ± 0.01	0.14 ± 0.01	0.07 ± 0.00
	VB-LR	0.68 ± 0.01	0.60 ± 0.03	$\mathbf{0.99 \pm 0.00}$	0.01 ± 0.01	0.13 ± 0.01	0.07 ± 0.01
	EB-LR	0.68 ± 0.01	0.61 ± 0.03	$\mathbf{0.99 \pm 0.00}$	0.01 ± 0.01	0.13 ± 0.01	0.07 ± 0.01
Idiographic	**RBF-GP**	0.84 ± 0.00	$\mathbf{0.73 \pm 0.01}$	0.97 ± 0.00	0.10 ± 0.03	0.53 ± 0.02	0.48 ± 0.02
	KNN	0.77 ± 0.01	0.72 ± 0.02	0.97 ± 0.00	0.09 ± 0.03	0.42 ± 0.02	0.39 ± 0.04
	LR	0.85 ± 0.01	$\mathbf{0.73 \pm 0.01}$	0.97 ± 0.00	0.09 ± 0.01	0.55 ± 0.02	0.47 ± 0.03
	VB-LR	$\mathbf{0.87 \pm 0.00}$	$\mathbf{0.73 \pm 0.01}$	0.96 ± 0.00	0.12 ± 0.03	0.57 ± 0.02	0.53 ± 0.02
	EB-LR	0.84 ± 0.01	0.71 ± 0.02	0.96 ± 0.00	0.27 ± 0.05	0.58 ± 0.02	0.54 ± 0.03
	SV-LSGP	0.82 ± 0.01	0.57 ± 0.03	0.91 ± 0.01	$\mathbf{0.29 \pm 0.07}$	$\mathbf{0.62 \pm 0.03}$	$\mathbf{0.57 \pm 0.03}$

Gaussians $q(z_n; \phi)$. This is equivalent to maximizing the evidence lower bound (ELBO) [19]:

$$\mathcal{L} = \sum_i \mathbb{E}_{q(f_i|X;W,\theta)} \left[\log p(y_i|f_i)\right] - D_{\mathrm{KL}}[q(U;\phi)||p(U;W,\theta)] - \sum_{n=1}^{N} D_{\mathrm{KL}}[q(z_n;\phi)||p(z)] \quad (2)$$

wherein the expectation is approximated via Monte Carlo by sampling $q(f_i|X;W,\theta) = \mathbb{E}_{q(U;\phi)}\left[p(f_i|U,X,Z;W,\theta)\right] \cdot \prod_{n=1}^{N} q(z_n;\phi)$, where the expectation is computed analytically [20]: $\mathcal{N}\left(\psi_i \cdot \mu_\phi, \mathrm{diag}(\psi_i \cdot (\Sigma_\phi - K_\theta(W,W)) \cdot \psi_i^\intercal)\right)$, with $\psi_i = K_\theta(\hat{x}_i^\intercal, W) \cdot K_\theta(W,W)^{-1}$. Since the first term of $\mathcal{L}$ can be estimated via mini-matching [20–22], performance is only dominated by $O(M^3)$ per gradient step.

Visualizing Latent Similarity. We can visualize the similarity between patients even in high dimensional latent spaces using a graph, provided that the kernel over $\hat{x}$ can be decomposed into a product of kernels applied to x and z. We compute the covariance matrix between *patients* (not observations) by applying the latent-space kernel to the variational means. We then treat this covariance as a graph adjacency matrix, in which in which every node is a patient and edge widths are proportional to the covariance between the patients. To reduce visual clutter, we prune edges with covariance below a chosen threshold. This can help us identify clusters of patients who borrow strength from one another and to explore how these patterns align with social determinants of health and other relevant factors to deepen our understanding of patient similarity.

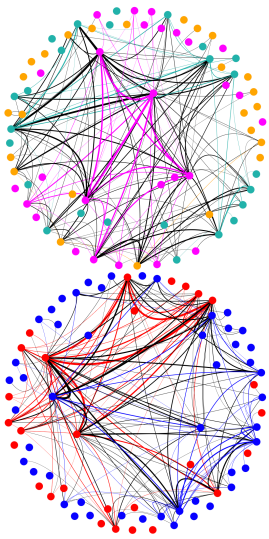


Figure 2: **Graph of Patient Similarity.** Nodes: adult adolescent, and top 33% middle 33% bottom 33%. Edges are black if connecting nodes of different colors; thickness indicates magnitude of covariance.

4 Experiments, Results, and Future Work

Preliminary results show that our method is not far from the best baselines on most metrics, outperforming all on sensitivity, which is crucial for suicide prevention. We compare our method with baselines in our ability to better forecast SREs one week in advance (details in Appendix B). Table 1 shows that, only having naively experimented with *a single kernel*, our method already matches the better performing methods on most metrics, obtaining worse PPV but better sensitivity. We anticipate that a future investigation into the inductive biases of different kernels will allow our method to outperform all baselines, since the LSGP generalizes the GP and LR methods.

Insights from Patient Similarity Graphs. We visualize the similarity of patients in Fig. 2. In the top, colors represent stratification of patients into bottom, middle, and top thirds based on SRE count. The figure shows that forecasts for patients with fewest SREs draw on each other and on patients with the most SREs (black edges connect teal and magenta). In the bottom of Fig. 2, color represents adult vs. adolescent, showing that, while we expect the trend for adults to differ from those of adolescents, the story may be more complicated. In future work, we hope to explore patient similarity based on other factors, such as social determinants of health.

References

[1] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

[2] Evan M Kleiman, Catherine R Glenn, and Richard T Liu. The use of advanced technology and statistical methods to predict and prevent suicide. *Nature reviews psychology*, 2(6):347–359, 2023.

[3] Evan M Kleiman, Brianna J Turner, Szymon Fedor, Eleanor E Beale, Jeff C Huffman, and Matthew K Nock. Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. *Journal of abnormal psychology*, 126(6):726, 2017.

[4] Ewa K Czyz, Cheryl A King, Nadia Al-Dajani, Lauren Zimmermann, Victor Hong, and Inbal Nahum-Shani. Ecological momentary assessments and passive sensing in the prediction of short-term suicidal ideation in young adults. *JAMA Network Open*, 6(8):e2328005–e2328005, 2023.

[5] Chang Lei, Diyang Qu, Kunxu Liu, and Runsen Chen. Ecological momentary assessment and machine learning for predicting suicidal ideation among sexual and gender minority individuals. *JAMA network open*, 6(9):e2333164–e2333164, 2023.

[6] Shirley B Wang, Ruben DI Van Genugten, Yaniv Yacoby, Weiwei Pan, Kate H Bentley, Suzanne A Bird, Ralph J Buonopane, Alexis Christie, Merryn Daniel, Dylan DeMarco, et al. Building personalized machine learning models using real-time monitoring data to predict idiographic suicidal thoughts. *Nature Mental Health*, pages 1–10, 2024.

[7] Kathryn R Fox, Xieyining Huang, Eleonora M Guzmán, Kensie M Funsch, Christine B Cha, Jessica D Ribeiro, and Joseph C Franklin. Interventions for suicide and self-injury: A meta-analysis of randomized controlled trials across nearly 50 years of research. *Psychological bulletin*, 146(12):1117, 2020.

[8] Ewa K Czyz, Cheryl A King, and Inbal Nahum-Shani. Ecological assessment of daily suicidal thoughts and attempts among suicidal teens after psychiatric hospitalization: Lessons about feasibility and acceptability. *Psychiatry research*, 267:566–574, 2018.

[9] Evan M Kleiman, Brianna J Turner, Szymon Fedor, Eleanor E Beale, Rosalind W Picard, Jeff C Huffman, and Matthew K Nock. Digital phenotyping of suicidal thoughts. *Depression and anxiety*, 35(7):601–608, 2018.

[10] Aleksandra Kaurin, Alexandre Y Dombrovski, Michael N Hallquist, and Aidan GC Wright. Integrating a functional view on suicide risk into idiographic statistical models. *Behaviour research and therapy*, 150: 104012, 2022.

[11] Daniel DL Coppersmith, Evan M Kleiman, Alexander J Millner, Shirley B Wang, Cara Arizmendi, Kate H Bentley, Dylan DeMarco, Rebecca G Fortgang, Kelly L Zuromski, Joseph S Maimone, et al. Heterogeneity in suicide risk: Evidence from personalized dynamic models. *Behaviour research and therapy*, 180:104574, 2024.

[12] Keith Hawton and Kees Van Heeringen. *The international handbook of suicide and attempted suicide*. John Wiley & Sons, 2000.

[13] Hilario Blasco-Fontecilla, Maria Rodrigo-Yanguas, Lucas Giner, Maria Jose Lobato-Rodriguez, and Jose De Leon. Patterns of comorbidity of suicide attempters: an update. *Current psychiatry reports*, 18(10):93, 2016.

[14] Chunyi Wang and Radford M Neal. Gaussian process regression with heteroscedastic or non-gaussian residuals. *arXiv preprint arXiv:1212.6246*, 2012.

[15] Kaspar Märtens, Kieran Campbell, and Christopher Yau. Decomposing feature-level variation with covariate Gaussian process latent variable models. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4372–4381. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/martens19a.html`.

[16] Didong Li, Andrew Jones, Sudipto Banerjee, and Barbara E. Engelhardt. Bayesian multi-group gaussian process models for heterogeneous group-structured data. *Journal of Machine Learning Research*, 26(30): 1–34, 2025. URL `http://jmlr.org/papers/v26/23-0291.html`.

[17] Steindór Sæmundsson, Katja Hofmann, and Marc Peter Deisenroth. Meta reinforcement learning with latent variable gaussian processes. *arXiv preprint arXiv:1803.07551*, 2018.

[18] Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR, 2009.

[19] Michalis Titsias and Neil D. Lawrence. Bayesian gaussian process latent variable model. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 844–851, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL `https://proceedings.mlr.press/v9/titsias10a.html`.

[20] James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable Variational Gaussian Process Classification. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 351–360, San Diego, California, USA, 09–12 May 2015. PMLR. URL `https://proceedings.mlr.press/v38/hensman15.html`.

[21] James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 282–290, Arlington, Virginia, USA, 2013. AUAI Press.

[22] Vidhi Lalchand, Aditya Ravuri, and Neil D. Lawrence. Generalised gplvm with stochastic variational inference. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7841–7864. PMLR, 28–30 Mar 2022. URL `https://proceedings.mlr.press/v151/lalchand22a.html`.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] Amazasp Shaumyan. GitHub: AmazaspShumik/sklearn-bayes: Python package for Bayesian Machine Learning with scikit-learn API. `https://github.com/AmazaspShumik/sklearn-bayes?tab=readme-ov-file#contributions`. [Accessed 01-09-2025].

[25] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1):973–978, 2019.

[26] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

A Overview of the EMA Data

Participants. A total of 638 participants presenting with suicidal thoughts and/or recent suicidal behavior were recruited from two hospitals in the Boston area—318 adults (ages 18+) from a psychiatric emergency service, and 320 adolescents (ages 12-19) from a psychiatric inpatient unit. Participants were excluded if they did not own an iOS/Android smartphone, they presented any factor that impaired their ability to provide informed consent/assent, an inability to speak or write English fluently, a gross cognitive impairment due to florid psychosis, intellectual disability, dementia, acute intoxication, or extremely agitated or violent behavior.

Consent, Compensation, and IRB. After agreeing to participate, individuals signed consent/assent forms, answered an initial questionnaire, and installed the LifeData application on their mobile devices, which prompted them with brief self-report questionnaires. Participants received $10 for completing the initial questionnaire and earned $1 for each EMA survey they submitted. The study was approved by our institutions' IRB.

Surveys. Smartphone surveys assessed participants' current experience of suicidal thinking—urge, intent, and ability to resist suicidal urges—as well as 17 affective states—negative, hopeless, trapped, isolated, burdensome, angry, self-hate, agitated, worried, numb, fatigued, humiliated, desire to escape, desire to avoid, energetic, and positive—on a 0-10 likert scale. These surveys were sent to participants 6-times per day for three months, with the first and last sent at fixed times decided in collaboration with each participant, and the remaining surveys sent at randomized times, at least two hours apart, and between the first and last surveys. In addition to these surveys, participants could always opt to

fill in additional surveys, for example, to report a suicide attempt, non-suicidal self-injury, or another event they deemed important. They study was monitored by a risk-monitoring team in real-time to intervene when participants indicate high suicidal intent (details available upon request).

Recording Suicide-Related Events (SREs). An SRE was recorded in the data if it was reported by the patient in the survey, if it was reported by the risk-monitoring team, or if it was reported in the patient's electronic health record (consensus coded by two trained BA-level reviewers with supervision by a doctoral-level clinician with expertise in assessing/treating STBs).

Data Inclusion in Analysis. We kept all SREs for which there was at least one EMA survey in the week prior. We kept data from all patients that had at least 3 SREs and 3 non-SREs to ensure we can include one of each in the train/validation/test split (see Appendix B). Due to the low base-rate of SREs, this left us with $N = 77$ patients who contributed a total of 14763 complete EMA surveys.

B Experimental Setup

Data Splits. We divided the data into 50%, 25%, and 25% sized-sets for training, validation, and test, respectively. We ensured that there was at least one SRE and one non-SRE in each set. As such, we assume that for our method to be used in practice, patients must have at least one recorded SRE in their data. We created these cuts of the data 5 times, conducting all experiments on each cut of the data, and reporting the mean $\pm$ standard deviation of all metrics.

Random Restarts. For each of cut of the data, we ran each method 5 times, each with a random seed. We selected the best performing random restart on the validation ROC-AUC.

Hyperparameter Selection. We performed grid search over the following parameters, selecting them based on ROC-AUC on the validation set:

- **KNN:** Neighbors $k \in \{1, 2\}$, which performed best in our preliminary experiments, and distance $\in \{\text{Minkowski}, \text{Manhattan}\}$. We used the default parameters from `scikit-learn` [23] for the remaining parameters.
- **LR:** Default parameters from `scikit-learn` [23] but with a maximum of 5000 iterations until convergence.
- **VB-LR:** We trained for a maximum of 5000 iterations until convergence, with the rate and scale α, β on the Gamma prior on precision of the coefficients both $\in \{1.0, 2.0\}$, and with the remaining parameters set to the defaults from Shaumyan [24].
- **EB-LR:** We trained for a maximum of 5000 iterations, with the initial precision of prior distribution $\alpha \in \{3.0, 2.0, 1.0, 1\text{e}{-3}, 1\text{e}{-6}, 1\text{e}{-9}, 1\text{e}{-12}\}$, and with the remaining parameters set to the defaults from Shaumyan [24].
- **GP:** We used the default GP hyperparameters from `scikit-learn` [23], which uses an automatic relevance determination (ARD) kernel. We additionally set `max_iter_predict` $= 5000$, as well as `n_restarts_optimizer` $= 1$, which selects across two kernel hyperparameter initializations—default and random.
- **SV-LSGP:** We use $M = 2000$ inducing points, $D_z = 3$, mini-batch size $B = 256$, and a kernel that factorizes as $K_\theta(\widehat{X}, \widehat{X}') = K^x_\theta(X, X') \cdot K^z_\theta(Z, Z')$, with K^x_θ as a linear kernel and K^z_θ as an arccos kernel. We fit the model with 60000 gradient steps and a learning rate of 0.001.

Software. We implemented the SV-LSGP in NumPyro [25] and Jax [26].