Modeling PTSD Trajectories with Conditional SVAEs and Synthetic Data Generation: Data-Efficient Prediction and Outcome-Specific Explainability

Mateus Guimarães Lima de Freitas
Department of Computer Science
University of Texas at Austin

Alexander Rasgon, MD
Dell Medical School
University of Texas at Austin

Shuangyu Li
Department of Computer Science
University of Texas at Austin

Zhan Chen
Department of Computer Science
University of Texas at Austin

Dongjin Song, PhD
Department of Computer Science and Engineering
University of Connecticut

João Paulo Abreu Maranhão, PhD
Systems Development Center
Brazilian Army

Yunyu Xiao, PhD
Department of Population Health Sciences
Weill Cornell Medicine

Ying Ding, PhD
School of Information
University of Texas at Austin

Abstract

Childhood trauma initiates complex psychiatric trajectories, but predictive modeling is hampered by data scarcity. We ask if a conditional Transformer-based Sequential Variational Autoencoder (SVAE) can learn patient embeddings from longitudinal surveys to improve Post-traumatic Stress Disorder (PTSD) prediction and reveal clinical drivers of model performance. Our framework uses a conditional SVAE to generate synthetic patient trajectories, addressing class imbalance. In our experiments, combining real and synthetic data **increased the identification of true positive PTSD cases by 82%** over a real-data-only baseline, achieving a top F1-score of 0.683. Ablation studies confirm that architectural choices like "free bits" are essential for generating effective augmentation data. Finally, by stratifying SHAP explanations by outcome (True Positives, False Positives, False Negatives, and True Negatives), we transform interpretability into a diagnostic tool, revealing how the model's reasoning differs between correct and incorrect predictions. This allows for targeted clinical insights, such as identifying when the model over-weights hopelessness signals, making predictions more transparent and clinically actionable.

1 Introduction

Childhood trauma, encompassing a range of adverse childhood experiences (ACEs), represents a profound public health crisis with enduring consequences for mental health [13]. The prevalence of such experiences is alarmingly high; a large-scale survey by the World Mental Health (WMH) Initiative found that nearly 40% of adults reported at least one ACE [20]. Studies focusing on

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

psychiatric populations reveal even higher rates, with a childhood trauma prevalence as high as 85% [24]. The etiological link between these early adversities and the subsequent development of psychiatric disorders is unequivocally established. Victims of childhood abuse exhibit staggering rates of subsequent psychopathology, with studies indicating that as many as 80% develop depression and over 50% suffer from anxiety [24, 36]. This connection is not limited to mood and anxiety disorders; childhood trauma is a significant risk factor for a wide spectrum of conditions, including PTSD, psychosis, personality disorders, and substance use disorders [20, 53]. This work addresses the following research question: **Can a conditional Transformer-SVAE trained on dense, longitudinal Clinical and Social Determinants of Health (SDoH) surveys learn patient-level trajectory embeddings that (i) improve PTSD trajectory prediction under small-N/class-imbalance and (ii) reveal outcome-specific drivers of model success and failure?** If effective, this approach would **(a) increase identification of at-risk youth trajectories** (more true positives with stable precision), and **(b) provide visit- and instrument-level explanations** that clinicians can act on (e.g., when hopelessness signals are over-weighted versus attenuated by protective factors).

However, a simple correlational link fails to capture the full complexity of trauma's impact. Early life adversity does not merely increase the probability of a future diagnosis; it fundamentally alters the **developmental trajectory** of psychopathology [13]. The experience of trauma can shape the age of onset, clinical severity, symptom presentation, and long-term course of mental illness [13]. This evidence reframes the central scientific challenge: it is not sufficient to predict a static diagnostic outcome. Rather, the goal must be to model the entire pathological process—the dynamic, evolving trajectory of mental health or illness as it unfolds over an individual's life course. This perspective necessitates a shift from conventional cross-sectional analysis towards sophisticated longitudinal modeling capable of capturing the nuances of disease progression over time [19, 37]. Compounding this challenge is the growing recognition that non-clinical factors, or SDoH, are powerful drivers of these trajectories. Recent large-scale analyses have demonstrated that multidimensional SDoH profiles—encompassing economic, social, and environmental factors—are strongly correlated with mental health outcomes such as suicide rates, with distinct regional and demographic patterns [51]. This evidence underscores the necessity of integrating comprehensive SDoH data into any predictive framework aiming for clinical and social relevance [51, 54]. Furthermore, analogous research in adolescent mental health has shown that the trajectory of risk factors, such as addictive screen use, is more predictive of adverse outcomes than static, single-time-point measurements, reinforcing the imperative to adopt a longitudinal perspective [50].

This imperative to model dynamic, SDoH-informed trajectories is met with a formidable data-methodology chasm. The requisite longitudinal datasets, while rich in temporal detail, are often characterized by high dimensionality, privacy constraints, and, most critically, small sample sizes, particularly for vulnerable populations and specific diagnoses like PTSD [27]. This data reality is in direct conflict with the requirements of the deep learning models best suited for the task. Architectures like the Transformer have shown unparalleled success in capturing long-range dependencies in sequential data, making them state-of-the-art for analyzing clinical time series [49, 27, 39]. However, these models are notoriously data-intensive, and their effectiveness is often hampered by the very data scarcity they are meant to address [8]. This fundamental mismatch necessitates a new approach, where generative data augmentation emerges as a critical and enabling methodology for bridging this gap [26, 34].

To address these challenges, this paper introduces a comprehensive framework for modeling and interpreting psychiatric trajectories using a longitudinal dataset of pediatric patients exposed to trauma. Our contribution is threefold. First, we develop a **Conditional Transformer-based Sequential Variational Autoencoder** specifically designed to learn holistic, dynamic embeddings from dense, multi-visit patient records. This architecture provides the foundation for a principled latent-space data augmentation strategy that generates realistic, synthetic patient trajectories to mitigate data scarcity and class imbalance [22, 45, 3]. Second, we incorporate key methodological enhancements to ensure the robustness of the generative model, including the **"free bits" technique** to prevent posterior collapse—a common failure mode in VAEs—and thereby promote the learning of meaningful latent representations [21, 17, 15, 12, 18]. Third, we introduce a novel **outcome-specific explainability analysis** using SHAP (SHapley Additive exPlanations). Moving beyond standard global feature importance, we systematically dissect the model's reasoning by comparing feature contributions across correct (True Positive) and incorrect (False Negative) predictions. This granular analysis transforms interpretability from a simple validation exercise into a powerful diagnostic tool for

discovering the clinical drivers of model success and failure, paving the way for more trustworthy and clinically actionable predictive systems in computational psychiatry [30, 32, 5].

The contributions of this framework directly map to clinical decision-making. By improving the identification of true positive cases through generative augmentation, our model can support **earlier escalation of care** for at-risk youth. Furthermore, the outcome-specific explanations provide a powerful diagnostic tool. For instance, when the SHAP analysis flags that high-potency items from the Concise Health Risk Tracking (CHRT) scale are spurious drivers of a False Positive prediction, it prompts a clinician to perform a more targeted assessment, preventing potential alarm fatigue. Finally, by identifying robust predictive signals from SDoH factors, this work provides quantitative evidence to guide the allocation of resources and inform public health policy.

2 Related Work

2.1 Longitudinal Trajectory Modeling in Psychiatry

Modeling the progression of psychiatric illness over time has historically relied on statistical methods such as linear mixed-effects models [37]. While valuable for identifying population-level trends, these methods often struggle with the high dimensionality, non-linearity, and complex interactions present in modern clinical datasets. The advent of deep learning introduced more powerful alternatives, with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks becoming standard choices for analyzing sequential health data [28]. More recently, Transformer-based architectures, with their self-attention mechanism, have demonstrated superior performance in capturing long-range dependencies within longitudinal data, including Electronic Health Records (EHRs), establishing them as the state-of-the-art for tasks like disease progression modeling and risk prediction [27, 39, 40]. Research in PTSD prediction specifically has begun to leverage machine learning, with systematic reviews indicating a prevalence of tree-based models and highlighting a critical need for more robust deep learning applications and rigorous external validation [2]. Other emerging approaches include the use of functional brain imaging to predict symptom trajectories [7] and graph induction models to infer potential causal pathways to illness from longitudinal data [37]. Our work contributes to this area by employing a Transformer-based architecture within a generative framework, specifically tailored to the dense, fixed-interval survey data common in structured clinical research.

2.2 Generative Models for Synthetic Health Data

Data scarcity, class imbalance, and privacy regulations are significant barriers in medical research [35, 26]. Deep generative models have emerged as a powerful solution for creating high-fidelity synthetic health data to address these challenges [34]. Generative Adversarial Networks (GANs), such as MedGAN, have been used to generate realistic patient records, but are often plagued by training instability [4]. Variational Autoencoders (VAEs) offer a more stable, probabilistic alternative that learns a well-structured latent space, making them particularly suitable for controlled data augmentation via sampling and interpolation [22, 45, 3]. A critical distinction in longitudinal modeling is the nature of the input data. Many state-of-the-art sequential VAEs for healthcare, such as IVP-VAE [38] or Shi-VAE [6], are designed for sparse, irregularly-sampled event sequences typical of EHR billing or lab data. These models excel at handling irregular timestamps and high degrees of missingness. In contrast, our work focuses on a different but equally important data modality: dense, fixed-interval sequences, where each visit yields a complete, high-dimensional vector of features from comprehensive surveys. For this data structure, a Sequence-to-Sequence (Seq2Seq) VAE architecture is more appropriate, as it is designed to learn a holistic representation of an entire sequence of dense vectors, a task for which event-based models are ill-suited.

2.3 Advanced Topics in Variational Autoencoders

One challenge in training VAEs is "posterior collapse" or "KL vanishing," where the model learns to ignore the latent variable z, causing the KL divergence term in the loss function to approach zero and rendering the generative process ineffective [10]. To combat this, several techniques have been developed. KL annealing, which gradually increases the weight of the KL term during training, is a common strategy [15]. A more targeted approach is the "free bits" technique, which modifies the KL

loss to ensure that each dimension of the latent space is forced to encode at least a minimum amount of information [21, 17]. By framing the use of "free bits" not as a minor implementation detail but as a necessary component for learning robust representations, our work aligns with best practices for stable VAE training [12, 18]. Furthermore, our model architecture is conditional. Conditional VAEs (CVAEs) extend the VAE framework by incorporating auxiliary information into the generative process, allowing for more controlled and targeted data synthesis [42]. In clinical contexts, CVAEs have been used to generate data conditioned on specific patient attributes or diagnostic labels, providing a powerful mechanism for generating class-balanced datasets or exploring counterfactual scenarios [31, 55].

2.4 Interpretability in Clinical Machine Learning

The adoption of complex "black box" models in high-stakes clinical environments is contingent upon their interpretability [11, 29, 14]. Explainable AI (XAI) methods are therefore critical for building trust, validating models, and enabling clinical utility. SHAP (SHapley Additive exPlanations), a game-theoretic approach that provides a unified framework for explaining the output of any machine learning model, has become a gold standard in the field [30]. Its application in healthcare has been shown to provide clinically relevant insights by attributing predictions to specific input features [32, 5]. However, the vast majority of current XAI applications in medicine and psychiatry focus on generating global feature importance plots as a means of model validation—confirming that the model has learned "sensible" patterns that align with existing clinical knowledge [52, 19, 53]. While this is a necessary first step, it fails to explain why a model might be systematically failing for certain patients or how its reasoning process differs between correct and incorrect predictions. A significant gap therefore exists in the application of XAI for the purpose of model diagnostics and granular error analysis in a clinical setting. Our work directly addresses this gap by introducing an outcome-specific SHAP analysis that stratifies explanations by prediction outcome (e.g., True Positive vs. False Negative). This approach reframes interpretability as a discovery tool, enabling a deeper understanding of the model's decision boundaries and failure modes.

3 Methodology

This study employed a systematic machine learning pipeline to develop and evaluate models for predicting the multi-visit trajectory of Post-Traumatic Stress Disorder (PTSD) in a pediatric cohort. The methodology is centered on learning a holistic representation of patient trajectories using a conditional SVAE, leveraging this model for data augmentation, and training downstream classifiers for both prediction and granular, outcome-specific interpretation. The high-level overview of the modeling is represented on diagram 1 and an overview of the specific conditional SVAE modeling is represented on diagram 2.

3.1 Data Source and Cohort

Data were sourced from the Texas Childhood Trauma Research Network (TX-CTRN), a longitudinal study assessing trauma, diagnostic criteria, and Social Determinants of Health (SDoH) across multiple visits (baseline, 1 month, 6 months, 1 year, 18 months and 2 year visits), collected through a suite of validated surveys and clinician-administered interviews.

Key instruments used in this study captured three domains: **(1) Diagnosis and Symptom Severity**, using tools like the MINI-KID for diagnosis, the CAPS-CA-5 for PTSD symptoms, and the PHQ-A for depression; **(2) Trauma Exposure**, documented with the TESI-C; and **(3) Social Determinants of Health (SDoH)**, assessed with questionnaires covering demographics, social bonds, and cultural identity. A complete list of all instruments is available in Appendix.

3.2 Data Preprocessing and Feature Selection

The raw dataset, comprising multi-visit records for approximately 3,700 participants, underwent a multi-stage preprocessing pipeline to ensure data quality, handle the longitudinal structure, and construct a robust feature set for modeling.

First, initial data cleaning involved removing unscheduled events and standardizing null value representations (e.g., '999'). A key step to address the longitudinal format was the replication of time-invariant data; surveys administered only at baseline, such as demographics and trauma history (e.g., TESI-C), were forward-filled to all subsequent visits for each patient. The dataset was then refined through an automated filtering process that included: removing columns with near-zero variance or high proportions (>99%) of missing values; harmonizing columns with numerical suffixes (e.g., 'col.1') or different languages (e.g., 'col_sp'); and de-duplicating patient visit records to resolve multiple entries for a single event by selecting the most complete record.

Next, the feature set was transformed for machine learning compatibility. Remaining missing numerical values were imputed with zero, and a one-hot encoding scheme was applied to all low-cardinality categorical and numeric features, converting them into a binary format. Concurrently, diagnostic label columns (e.g., from MINI-KID and CAPS instruments) were separated from the features and binarized, with text-based entries ('Checked', 'Yes') converted to a '1/0' format. Discrepancies in diagnoses from multiple raters were reconciled by taking the maximum value, thereby creating a single, definitive label for each diagnosis per visit.

Finally, a rigorous feature selection protocol was executed to derive the most predictive and parsimonious feature set. To prevent data leakage, the cohort was first partitioned into training (80%) and test (20%) sets using a strict patient-level split, ensuring all records for any given patient belonged to only one set [48]. An empirical, performance-based selection process was then conducted exclusively on the training data. A `RandomForestClassifier` was used to rank all features by Gini importance. We then systematically evaluated the performance of this classifier on several feature subsets using 5-fold grouped cross-validation to respect the data's patient-visit structure. The analysis revealed a clear performance peak; the classifier achieved its maximum cross-validated ROC-AUC using the top-ranked features that accounted for just 35% of the total cumulative Gini importance for the Clinical + SDoH data and 25% for the SDoH only data, accordingly to Figures 3 and 4. Including features beyond this point was found to be detrimental, indicating that this subset captured the strongest predictive signals while excluding features that introduced more noise than useful information.

3.3 Longitudinal Data Structuring and Splitting

To prepare the data for sequential modeling, records for each patient were grouped and ordered by visit date, creating a three-dimensional data structure of (patients $\times$ visits $\times$ features). To handle variable visit attendance and attrition, all patient sequences were padded to a uniform length of six visits (baseline, 1 month, 6 months, 1 year, 18 months and 2 years), with missing visits imputed using a Last Observation Carried Forward (LOCF) strategy. A binary mask was concurrently created to distinguish real from imputed data points. To ensure a robust and unbiased evaluation, a single, globally held-out test set was created via a stratified, patient-level split (80% train, 20% test) before any model training commenced. Stratification was performed based on a summary label representing each patient's overall PTSD trajectory (the MINI-KID diagnosis label), guaranteeing that the test set was both entirely unseen during training and representative of the cohort's diagnostic distribution.

3.4 The Conditional SVAE Framework for Trajectory Modeling

We employed a Conditional Sequential Variational Autoencoder (SVAE) as a powerful non-linear feature extractor for entire patient trajectories. The model is trained on a multi-task objective to simultaneously reconstruct the input trajectory $\mathbf{X}$ and predict a corresponding binary label y. This is achieved by minimizing a composite loss function that extends the standard Evidence Lower Bound (ELBO) [22]. The loss is defined as:

$$\mathcal{L}(\theta, \phi; \mathbf{X}, y) = w_{recon}\mathcal{L}_{recon} + w_{pred}\mathcal{L}_{pred} + w_{KL}D_{KL}(q_\phi(\mathbf{z}|\mathbf{X})\|p(\mathbf{z})) \quad (1)$$

where $\mathcal{L}_{recon}$ is the reconstruction loss (Mean Squared Error), $\mathcal{L}_{pred}$ is the prediction loss for the label (Binary Cross-Entropy), and the final term is the Kullback-Leibler (KL) divergence, which regularizes the latent space. The terms w_{recon}, w_{pred}, and w_{KL} are scalar weights for each component. The KL divergence encourages the approximate posterior $q_\phi(\mathbf{z}|\mathbf{X})$, learned from the data, to match a prior distribution $p(\mathbf{z})$, typically a standard normal $\mathcal{N}(0, \mathbf{I})$.

Our architecture consists of a Transformer-based encoder that maps a patient's multi-visit sequence $\mathbf{X}$ into the parameters of the approximate posterior $(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2)$, and a Transformer-based decoder that reconstructs the sequence $\mathbf{X}$ from a sample $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{X})$. In this methodology, conditionality

is implemented by embedding the ground-truth training `labels` (in our case the MINI-KID multi-diagnosis labels) and feeding them directly into the **decoder** alongside the latent vector `z`. The **encoder**, however, generates `z` using only the input features `src`, without ever seeing the labels. This design is sound for feature reconstruction ($p(x|z, y)$) but introduces data leakage if that same decoder is also asked to predict the labels it just received as input. Crucially, when this architecture is used in a two-stage process where the trained encoder first generates `z` from `src` and a separate classifier then predicts `y` from `z`—there is **no data leakage**. This is because the feature-generation step (`encode(src)`) is completely isolated from the labels, and this process benefits from the CVAE's ability to force `z` into a more disentangled and powerful representation of intra-class variation, which can improve the performance of the final classifier. This allows the generative process to be guided by specific patient characteristics and diagnoses, a technique well-established in the CVAE literature [42, 31, 55].

To ensure stable training and prevent posterior collapse, we implemented two key techniques. First, KL annealing gradually increases the weight of the KL divergence term (w_{KL}) during training [10]. Second, we employ a "free bits" threshold λ, which modifies the KL term to $\max(\lambda, D_{KL}(q_\phi(\mathbf{z}|\mathbf{X})\|p(\mathbf{z})))$, forcing the entire latent vector to encode a minimum amount of information and promoting the learning of richer representations [21, 17].

3.5 Latent Space Augmentation and Trajectory Prediction

Using the trained SVAE, we addressed data scarcity and class imbalance via class-conditional latent space augmentation. All real training trajectories were encoded into their latent vectors $\mathbf{z}^*_{\text{real}}$. These vectors were then separated by their PTSD summary label (positive vs. negative), and a multivariate Gaussian distribution ($\mathcal{N}(\boldsymbol{\mu}^*_{\text{pos}}, \boldsymbol{\Sigma}^*_{\text{pos}})$ and $\mathcal{N}(\boldsymbol{\mu}^*_{\text{neg}}, \boldsymbol{\Sigma}^*_{\text{neg}})$) was fitted to each class. New latent vectors were generated by sampling from these learned distributions and then passed through the trained decoder to generate complete, high-fidelity synthetic patient trajectories of the desired class. A suite of classifiers, including XGBoost and feed-forward neural networks, were then trained on the dynamic embeddings ($\mathbf{z}$) from the combined (real + synthetic) dataset to predict the 6-visit PTSD diagnostic trajectory. The standard threshold of 0.5 was used for classification decision. Hyperparameters for all models were systematically tuned using **Optuna**, an automated optimization framework, to maximize PR-AUC [1, 33, 25]. Detailed hyperparameter search ranges are provided in the Appendix.

3.6 Outcome-Specific Interpretability Pipeline

To move beyond model validation and toward clinical discovery, we implemented an end-to-end, outcome-specific explainability pipeline using `shap.KernelExplainer` to analyze the composite model, which chains the SVAE encoder with the final classifier [30]. This approach calculates SHAP values that trace a prediction for a specific visit back to the importance of each of the original features from any prior or concurrent visit in the patient's history. Critically, this analysis was stratified by the prediction outcome. SHAP values were aggregated and analyzed separately for four distinct groups: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). This stratification allows for a direct comparison of the features driving correct predictions versus those contributing to model errors, providing a granular view of the model's reasoning process and potential biases [32, 5].

4 Results

4.1 Predictive Performance with Generative Data Augmentation

Our experimental pipeline was replicated for two distinct feature subsets: one containing only SDoH features, and another containing all features from the TX-CTRN dataset. When restricted to SDoH features, the models demonstrated a clear benefit from data augmentation. As shown in Table 1, the Real Only scenario established a strong baseline, with XGBoost achieving the best F1-score of 0.660. The Synthetic Only scenario served as a crucial validation, with the XGBoost model reaching an F1-score of 0.622, confirming the VAE's ability to generate data with a valid predictive signal. The primary finding is in the Combined scenario, where the RandomForest model achieved the overall highest F1-score of 0.671. This improvement was driven by a substantial **41% increase** in the model's true positive count for PTSD cases (from 164 to 232) compared to the Real Only baseline.

The inclusion of rich clinical survey data significantly elevated the performance of all models. The detailed results in Table 2 show that the Real Only XGBoost model set a higher baseline with an F1-score of 0.671. The quality of the generative model in this richer feature space was also strong; classifiers trained on Synthetic Only data were highly competitive, with the Transformer model reaching an F1-score of 0.603. Once again, the benefit of augmentation was confirmed in the Combined scenario. The XGBoost model again achieved the top performance with an F1-score of **0.683**, driven by a dramatic **82% increase** in correctly identified PTSD cases (TP rose from 177 to 322).

Table 1: Detailed Model Performance (Visit-Level) for PTSD Prediction using SDoH Features Only.

Training Scenario	Model	ROC-AUC	PR-AUC	F1-Score
Real Only	XGBoost	0.798	0.463	**0.660**
	RandomForest	**0.800**	**0.469**	0.628
	FFN	0.752	0.409	0.558
	LSTM	0.743	0.417	0.569
	Transformer	0.747	0.407	0.537
Synthetic Only	XGBoost	0.767	0.411	**0.622**
	RandomForest	0.765	0.416	0.612
	FFN	0.751	0.418	0.571
	LSTM	0.752	0.425	0.595
	Transformer	**0.758**	**0.435**	0.605
Combined	XGBoost	**0.791**	**0.465**	**0.671**
	RandomForest	0.786	0.461	**0.671**
	FFN	0.759	0.427	0.568
	LSTM	0.753	0.431	0.590
	Transformer	0.758	0.435	0.622

Table 2: Detailed Model Performance (Visit-Level) using the Full Feature Set (SDoH + Clinical).

Training Scenario	Model	ROC-AUC	PR-AUC	F1-Score
Real Only	XGBoost	0.817	0.503	**0.671**
	RandomForest	0.821	0.504	0.664
	FFN	**0.831**	0.510	0.632
	LSTM	0.830	**0.512**	0.659
	Transformer	0.808	0.476	0.654
Synthetic Only	XGBoost	0.786	0.409	0.600
	RandomForest	**0.790**	0.423	0.602
	FFN	0.787	0.424	0.562
	LSTM	**0.790**	0.430	0.564
	Transformer	0.789	**0.434**	**0.603**
Combined	XGBoost	**0.819**	**0.476**	**0.683**
	RandomForest	0.816	0.472	0.674
	FFN	0.792	0.424	0.631
	LSTM	0.797	0.439	0.564
	Transformer	0.797	0.429	0.612

4.2 Ablation Study of "Free Bits" in Augmentation

To empirically validate our methodological choices, we conducted ablation studies quantifying the impact of key architectural components. The results reveal a nuanced but critical role for the **"free bits"** technique.

When evaluating the quality of synthetic data *in isolation*, the model trained *without* free bits held a slight edge, achieving a best PR-AUC of 0.441 compared to 0.434 from the full model. However, this

was reversed in the more practical **data augmentation** scenario. When synthetic data was combined with real data, the classifier trained on data from the "free bits" model was clearly superior, achieving a best PR-AUC of **0.476** compared to 0.456 from the non-free-bits model.

Most critically, this improvement in PR-AUC translated to a substantial clinical benefit. The classifier augmented with "free bits" data correctly identified **322 TPs** cases, a dramatic increase over the 287 cases found by the model augmented with non-free-bits data. This shows that while "free bits" might not produce the highest-fidelity data in isolation, it is essential for generating complementary synthetic samples that lead to a more robust and clinically effective final model. The benefit of the conditional architecture was also confirmed, as the unconditional SVAE performed worse than the full model across all metrics (Table 3).

Table 3: Ablation Study of SVAE Components. PR-AUC is reported for the best classifier in the *Combined* scenario.

Model Configuration	SVAE Val. Loss	PR-AUC (SDoH)	PR-AUC (Full)
Full Model (Conditional + Free Bits)	1.523	**0.465**	**0.476**
SVAE - No Free Bits	0.476	0.456	0.456
SVAE - Unconditional	—	0.369	0.361

4.3 Privacy Evaluation: Distance to Closest Record

To assess the privacy-preserving qualities of the generated synthetic data, we employ the Distance to Closest Record (DCR) metric [23, 16]. This evaluation is crucial to ensure that the synthetic trajectories are not mere copies of the original data, thereby mitigating re-identification risks.

The DCR score is calculated by first transforming each real and synthetic patient trajectory into a fixed-length feature vector using the SVAE's learned encoder ($\mathbf{z}$). For each synthetic trajectory, we then compute the Euclidean distance to every trajectory in the real training set and identify the minimum distance. The DCR is reported as the average of these minimum distances across all synthetic samples, with results shown in Table 4. The distance between the real training and test sets is included as a baseline to represent natural data variation.

Table 4: Distance to Closest Record (DCR) Scores. Lower scores indicate higher similarity to the real training data distribution.

Model Configuration	Feature Set	DCR (Synthetic vs. Train)	Baseline DCR (Test vs. Train)
Full Model (with Free Bits)	SDoH Only	0.01275	0.00734
Full Model (with Free Bits)	Clinical + SDoH	0.01181	0.00798
SVAE - No Free Bits	Clinical + SDoH	0.03302	0.02733

As shown in the table, the synthetic data generated by our **Full Model** has a DCR score that is remarkably close to the baseline (the natural distance between the real test and train sets). This indicates that the SVAE is generating high-fidelity samples that follow the true data distribution without merely replicating it.

Crucially, these results reinforce the findings from our ablation study. The model trained **without "free bits"** yields a significantly higher DCR (0.03302), suggesting it produces samples that are less faithful to the original data distribution. This aligns with its poorer performance in the data augmentation scenario and empirically confirms that the "free bits" technique is essential for generating higher-quality, more effective synthetic data.

4.4 Outcome-Specific Explainability via SHAP Analysis

To dissect the models' decision-making, we conducted an outcome-specific SHAP analysis. The top 20 mean SHAP values for each outcome for the best model overall (XGB) are presented on Figures 6, 8, 7 and 5.

In the models trained on the complete dataset, features from the **Concise Health Risk Tracking-Self Report (CHRT-16)** and the **Patient Health Questionnaire for Adolescents (PHQ-A)** were the most influential. For **TP**, the models correctly identified at-risk individuals by keying on expressions of hopelessness and functional impairment. The top-ranked features were consistently items like `chrt_3` ("It seems as if I can do nothing right") and `phqa_11` (related to difficulty in work, home, or social life).

Critically, the analysis of **FPs** revealed that the models were most often misled by the same instruments. Incorrect high-risk predictions were frequently driven by `chrt_7` ("I wish my suffering could just all be over") and other high-severity responses on the PHQ-A. This indicates a model vulnerability where expressions of significant distress are treated as definitive markers of risk, even when the full clinical picture may not support that conclusion. For **FNs**, the models failed to act on weaker but important signals, often from PHQ-A items related to anhedonia (`phq_2`) and low self-worth (`phq_6`), suggesting these features alone were insufficient to overcome the prediction threshold.

The models trained exclusively on the SDoH feature set adapted by identifying new proxies for risk from a different set of instruments. For **TPs**, the leading predictors were now items related to anxiety from the **Screen for Child Anxiety Related Disorders (SCARED)**, such as `scared_child_21` ("I worry about things working out for me"), and items from the PHQ-A concerning concentration (`phq_7`). The error patterns in this context shifted significantly. The leading drivers for **False Positives** were now features related to generalized anxiety and panic symptoms from the **SCARED** questionnaire (e.g., `scared_child_24`, "I get really frightened for no reason at all"). This demonstrates that when more direct clinical markers are unavailable, models may incorrectly interpret high anxiety as sufficient evidence of risk for the specific outcome being predicted, even if the two are not directly correlated in a given patient.

5 Discussion and Conclusion

This study introduces and validates a comprehensive framework demonstrating how to build trustworthy and effective models for psychiatric trajectory prediction by coupling generative model enhancements with outcome-specific interpretability. Our contributions are threefold. First, our generative approach of combining real and synthetic data measurably improves prediction, increasing the number of correctly identified PTSD cases by a dramatic 82% in the full feature set compared to the real-only baseline without collapsing precision. The strong performance of classifiers trained on synthetic data alone validates that our SVAE captured the complex, underlying distribution of the real data. Second, our ablation studies empirically justify key architectural choices, showing that techniques like "free bits" generate more complementary synthetic samples, which in turn improves downstream PR-AUC when mixed with real data. Third, our outcome-specific explainability analysis moves XAI from a simple validation exercise to a powerful diagnostic tool for error analysis. By stratifying SHAP values, we identified precisely why the model fails, such as its over-reliance on high-potency items from the CHRT and PHQ-A questionnaires for False Positives and its under-weighting of anhedonia signals for FNs.

A key insight is the double-edged nature of dominant predictors from certain clinical instruments. The models' reliance on high-potency items from instruments measuring **hopelessness (CHRT-16)** and **depression (PHQ-A)** is logical and effective for identifying TPs. However, the consistent appearance of the same features as top drivers for False Positives highlights a critical failure in contextual reasoning. This suggests the models employ a heuristic—"expressed hopelessness equals imminent risk"—without adequately weighing other protective or nuanced factors present in the data. This has significant clinical implications, as it could lead to alarm fatigue or the misinterpretation of a patient's existential distress if not reviewed carefully by a human expert. Furthermore, the analysis of the SDOH-only models uncovers the models' strategy of **proxy-based reasoning**. When primary depressive and hopelessness signals are less prominent, the models pivot to concepts of **anxiety**, as measured by the **SCARED** instrument, as a stand-in for clinical risk. This is a powerful demonstration of the models' ability to find signals in data that might be considered secondary in a traditional clinical assessment. However, this strategy also introduces a distinct bias: the tendency to conflate generalized anxiety with the specific psychiatric outcome. This suggests a risk of penalizing patients for traits or symptoms of anxiety, which, while indicative of suffering, may not map directly to the predicted risk. This distinction is vital for building fair and clinically useful predictive tools that understand the multifaceted nature of psychiatric symptoms.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[2] Mohammed Al-Hashedi et al. Machine learning algorithms for predicting ptsd: a systematic review and meta-analysis. *BMC Medical Informatics and Decision Making*, 25(1):1–15, 2025.

[3] Muhammad Aqeel, Maham Nazir, Zanxi Ruan, and Francesco Setti. Latent space synergy: Text-guided data augmentation for direct diffusion biomedical segmentation. *arXiv preprint arXiv:2507.15361*, 2025.

[4] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized Medical Imaging and Graphics*, 79:101684, January 2020.

[5] Ponce Bobadilla AV, Schmitt V, Maier CS, Mensing S, and Stodtmann S. A practical guide to shapley additive explanations (shap) for ml model interpretability in computational biology. *Clin Transl Sci. 2024*, 2024.

[6] Daniel Barrejon, Pablo M Olmos, and Antonio Artes-Rodriguez. Medical data wrangling with sequential variational autoencoders. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2737–2745, 2022.

[7] Ziv Ben-Zion et al. Machine learning model predicts ptsd symptom severity over time. *JAMA Network Open*, 2025. As reported in Yale News, March 10, 2025.

[8] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[9] B. Birmaher, S. Khetarpal, M. Cully, D. Brent, and S. McKenzie. Screen for child anxiety related emotional disorders (scared) child & parent versions. [Measurement instrument], 1997.

[10] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, 2016.

[11] T Chen et al. Interpretable deep learning framework for nonlinear signal processing in intelligent psychiatric diagnostics. *Frontiers in Psychiatry*, 2025.

[12] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

[13] Andrea Danese and Melvyn Tan. Childhood maltreatment and obesity: systematic review and meta-analysis. *Molecular psychiatry*, 19(5):544–554, 2009.

[14] Koutsouleris N. Dwyer DB, Falkai P. Machine learning approaches for clinical psychology and psychiatry. *Annu Rev Clin Psychol.*, 2018.

[15] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 240–250, 2019.

[16] Chufan Gao, Mandis Beigi, Afrah Shafquat, Jacob Aptekar, and Jimeng Sun. Trialsynth: Generation of synthetic sequential clinical trial data. In *GenAI for Health: Potential, Trust and Policy Compliance*, 2024.

[17] Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *International Conference on Learning Representations*, 2019.

[18] Rasmus Malik Thaarup Høegh. Representation learning with variational autoencoders. 2022.

[19] Dan W. Joyce, Andrey Kormilitzin, Katharine A. Smith, and Andrea Cipriani. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *NPJ Digital Medicine*, 6:6, 2023.

[20] Ronald C Kessler, Katie A McLaughlin, Jennifer G Green, Michael J Gruber, Nancy A Sampson, Alan M Zaslavsky, Sergio Aguilar-Gaxiola, Ali Obaid Alhamzawi, Jordi Alonso, Matthias Angermeyer, et al. Childhood adversities and adult psychopathology in the who world mental health surveys. *The British Journal of Psychiatry*, 197(5):378–385, 2010.

[21] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016.

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[23] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17564–17579. PMLR, 23–29 Jul 2023.

[24] Sireesha Kothapalli, Asit Kumar Mishra, Snehil Guha, and OP Singh. Prevalence of childhood trauma in psychiatric patients: A cross-sectional study. *Industrial Psychiatry Journal*, 33(1):155, 2024.

[25] Li-Hsing Lai, Ying-Lei Lin, Yu-Hui Liu, Jung-Pin Lai, Wen-Chieh Yang, Hung-Pin Hou, and Ping-Feng Pai. The use of machine learning models with optuna in disease prediction. *Electronics*, 13(23), 2024.

[26] Jin Li, Benjamin J. Cairns, Jingsong Li, and Tingting Zhu. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *npj Digital Medicine*, 6:98, 2023.

[27] Y Li et al. Transformers in health: a systematic review on architectures for longitudinal data analysis. *Artificial Intelligence Review*, pages 1–39, 2024.

[28] Zachary C. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning, 2015.

[29] Xiaozheng Liu, Zhi Xu, Weikai Li, Zhen Zhou, and Xize Jia. Editorial: Clinical application of machine learning methods in psychiatric disorders. *Frontiers in Psychiatry*, Volume 14 - 2023, 2023.

[30] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[31] Joseph Masci et al. Leveraging conditional variational autoencoders for functional connectivity analysis in autism spectrum disorder. *Applied Sciences*, 13(23):12699, 2023.

[32] Yiannis Mastoras. Using shap to explain predictions in healthcare ml models. *Medium*, May 2024.

[33] Optuna Development Team. Optuna: A hyperparameter optimization framework. `https://optuna.readthedocs.io/`, 2024.

[34] Vasileios C. Pezoulas, Dimitrios I. Zaridis, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S. Tachos, and Dimitrios I. Fotiadis. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal*, 23:2892–2910, 2024.

[35] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[36] Angelo Sadeghpour, Varsha D. Badal, David L. Pogge, Elizabeth O'Donoghue, Tim Bigdeli, and Philip D. Harvey. Using machine learning modeling to identify childhood abuse victims on the basis of personality inventory responses. *Journal of Psychiatric Research*, 180:8–15, 2024.

[37] Glenn N. Saxe, Sisi Ma, Jiwen Ren, and Constantin F. Aliferis. Machine learning methods to predict child posttraumatic stress: a proof of concept study. *BMC Psychiatry*, 17:223, 2017.

[38] M Schirmer et al. Ivp-vae: Modeling ehr time series with initial value problem solvers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12135–12143, 2023.

[39] Yijun Shao, Yan Cheng, Stuart J Nelson, Peter Kokkinos, Edward Y Zamrini, Ali Ahmed, and Qing Zeng-Treitler. Hybrid value-aware transformer architecture for joint learning from longitudinal and non-longitudinal clinical data. *medRxiv*, 2023.

[40] Yijun Shao, Yan Cheng, Stuart J Nelson, Peter Kokkinos, Edward Y Zamrini, Ali Ahmed, and Qing Zeng-Treitler. Hybrid value-aware transformer architecture for joint learning from longitudinal and non-longitudinal clinical data. *Journal of the American Medical Informatics Association*, 30(8):1433–1440, 2023.

[41] D. V. Sheehan, K. H. Sheehan, R. D. Shytle, J. Janavs, B. Bunting, and T. Hergueta. Mini-international neuropsychiatric interview for children and adolescents (mini-kid). [Measurement instrument], 2009.

[42] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[43] Texas Health and Human Services. Patient health questionnaire for adolescents (phq-a). [Measurement instrument], 2005.

[44] M. H. Trivedi, S. R. Wisniewski, D. W. Morris, M. Fava, J. K. Gollan, D. Warden, A. A. Nierenberg, B. N. Gaynes, M. M. Husain, J. F. Luther, S. Zisook, and A. J. Rush. Concise health risk tracking scale: A brief self-report and clinician rating of suicidal risk. *Journal of Clinical Psychiatry*, 72(6):757–764, 2011.

[45] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7064–7073, 2017.

[46] U.S. Department of Veterans Affairs, National Center for PTSD. Traumatic events screening inventory–child version (tesi-c). [Measurement instrument], 2016. Retrieved July 3, 2025, from https://www.ptsd.va.gov/professional/assessment/documents/TESI-C.pdf.

[47] U.S. Department of Veterans Affairs, National Center for PTSD. Clinician-administered ptsd scale for dsm-5 – child/adolescent version (caps-ca-5). [Measurement instrument], 2017.

[48] Andrius Vabalas, Eirini Gowen, Ellen Poliakoff, and Alexander J Casson. Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11):e0224365, 2019.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[50] Yunyu Xiao, Yuan Meng, Timothy T Brown, Katherine M Keyes, and J John Mann. Addictive screen use trajectories and suicidal behaviors, suicidal ideation, and mental health in us youths. *JAMA*, 2025.

[51] Yunyu Xiao, Yuan Meng, Timothy T Brown, Alexander C Tsai, Lonnie R Snowden, Julian Chun-Chung Chow, Jyotishman Pathak, and J John Mann. Machine learning to investigate policy-relevant social determinants of health and suicide rates in the united states. *Nature Mental Health*, 3(6):675–684, 2025.

[52] Li Zheng, Yu-juan Xue, Zhen-nan Yuan, and Xue-zhong Xing. Explainable shap-xgboost models for pressure injuries among patients requiring mechanical ventilation in intensive care unit. *Scientific Reports*, 15:9878, 2025.

[53] Xiaoxiao Zhou, Zongbao Liang, and Guangzhen Zhang. Using explainable machine learning to investigate the relationship between childhood maltreatment, positive psychological traits, and cptsd symptoms. *European Journal of Psychotraumatology*, 16:2455800, 2025.

[54] Y Zhou et al. Deep learning and large language models for social determinants of health extraction from clinical text. *arXiv preprint arXiv:2505.04655*, 2025.

[55] Y Zou et al. A conditional information bottleneck approach for time series imputation. *Advances in Neural Information Processing Systems*, 36, 2023.

A Appendix

A.1 Clinical Instruments and Surveys

The data for this study were collected from a comprehensive suite of validated clinical instruments and surveys. Table 5 lists the key instruments used to capture psychiatric diagnoses, symptom severity, trauma exposure, and Social Determinants of Health (SDoH).

Table 5: Key Clinical Instruments and Surveys Used for Data Collection.

Domain	**Instrument Name**	**Citation**
Diagnosis & Symptom Severity	Mini-International Neuropsychiatric Interview for Children and Adolescents (MINI-KID)	[41]
	Clinician-Administered PTSD Scale for DSM-5 – Child/Adolescent Version (CAPS-CA-5)	[47]
	Patient Health Questionnaire for Adolescents (PHQ-A)	[43]
	Screen for Child Anxiety Related Disorders (SCARED-C)	[9]
	Concise Health Risk Tracking Scale (CHRT)	[44]
Trauma Exposure	Traumatic Events Screening Inventory–Child Version (TESI-C)	[46]
Social Determinants of Health (SDoH)	Personal & Family History Questionnaire	N/A
	Inventory of Parent and Peer Attachment–Revised (IPPA-R)	N/A
	Ethnic Identity Scale-Brief (EIS-B)	N/A

A.2 Hyperparameter Search Ranges

Table 6 details the search spaces used for hyperparameter optimization with Optuna for both the Conditional Transformer SVAE and the downstream classifiers. The optimization was performed by maximizing the Area Under the Precision-Recall Curve (PR-AUC) on a validation set for the classifiers, and minimizing validation loss for the SVAE.

Table 6: Hyperparameter Search Ranges for Optuna Studies.

Model	**Hyperparameter**	**Range**	**Distribution/Type**
Conditional SVAE	Latent Dimension	{64, 128}	Categorical
	Transformer d_{model}	{128, 256}	Categorical
	Encoder Layers	[2, 4]	Integer
	Decoder Layers	[1, 3]	Integer
	Attention Heads	{2, 4, 8}	Categorical
	Learning Rate	$[10^{-5}, 10^{-3}]$	Log-Uniform
XGBoost	Num. Estimators	[200, 1000]	Integer (Step 100)
	Learning Rate	$[10^{-3}, 10^{-1}]$	Log-Uniform
	Max Depth	[3, 10]	Integer
	Subsample	[0.6, 1.0]	Uniform
	Colsample by Tree	[0.6, 1.0]	Uniform
FFN	Learning Rate	$[10^{-5}, 10^{-3}]$	Log-Uniform
	Batch Size	{32, 64, 128}	Categorical
	Dropout Rate	[0.1, 0.5]	Uniform
	Early Stopping Patience	[10, 30]	Integer
	Hidden Size	{128, 256}	Categorical
	Num. Layers	[2, 4]	Integer
LSTM	Learning Rate	$[10^{-5}, 10^{-3}]$	Log-Uniform
	Batch Size	{32, 64, 128}	Categorical
	Dropout Rate	[0.1, 0.5]	Uniform
	Early Stopping Patience	[10, 30]	Integer
	Hidden Size	{64, 128}	Categorical
	Num. Layers	[1, 3]	Integer
Transformer	Learning Rate	$[10^{-5}, 10^{-3}]$	Log-Uniform
	Batch Size	{32, 64, 128}	Categorical
	Dropout Rate	[0.1, 0.5]	Uniform
	Early Stopping Patience	[10, 30]	Integer
	d_{model}	{64, 128}	Categorical
	Attention Heads (n_{head})	{2, 4}	Categorical

B Supplementary Figures

B.1 Modeling Pipeline

Figures 1 and 2 illustrate the overall machine learning pipeline and the specific four-stage research methodology employed in this study, respectively.

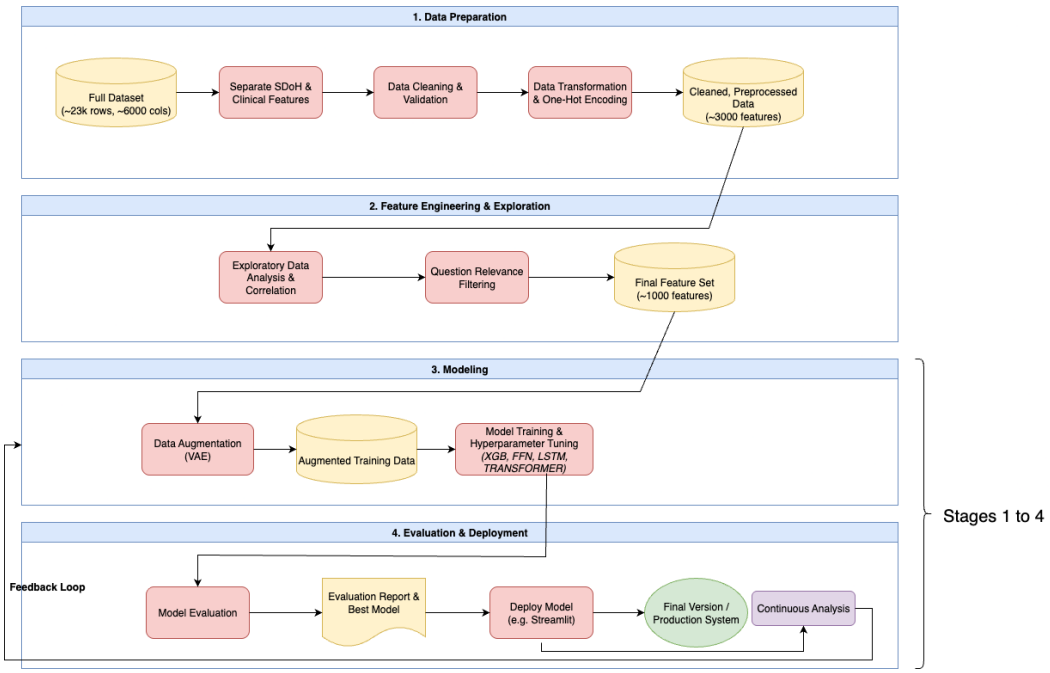

Figure 1: Preprocessing and Machine Learning Pipeline. This figure illustrates the comprehensive workflow, from initial data acquisition to final analysis. Key stages include data cleaning, transformation, filtering, VAE-based augmentation, model training, hyperparameter tuning, evaluation, and interpretation.

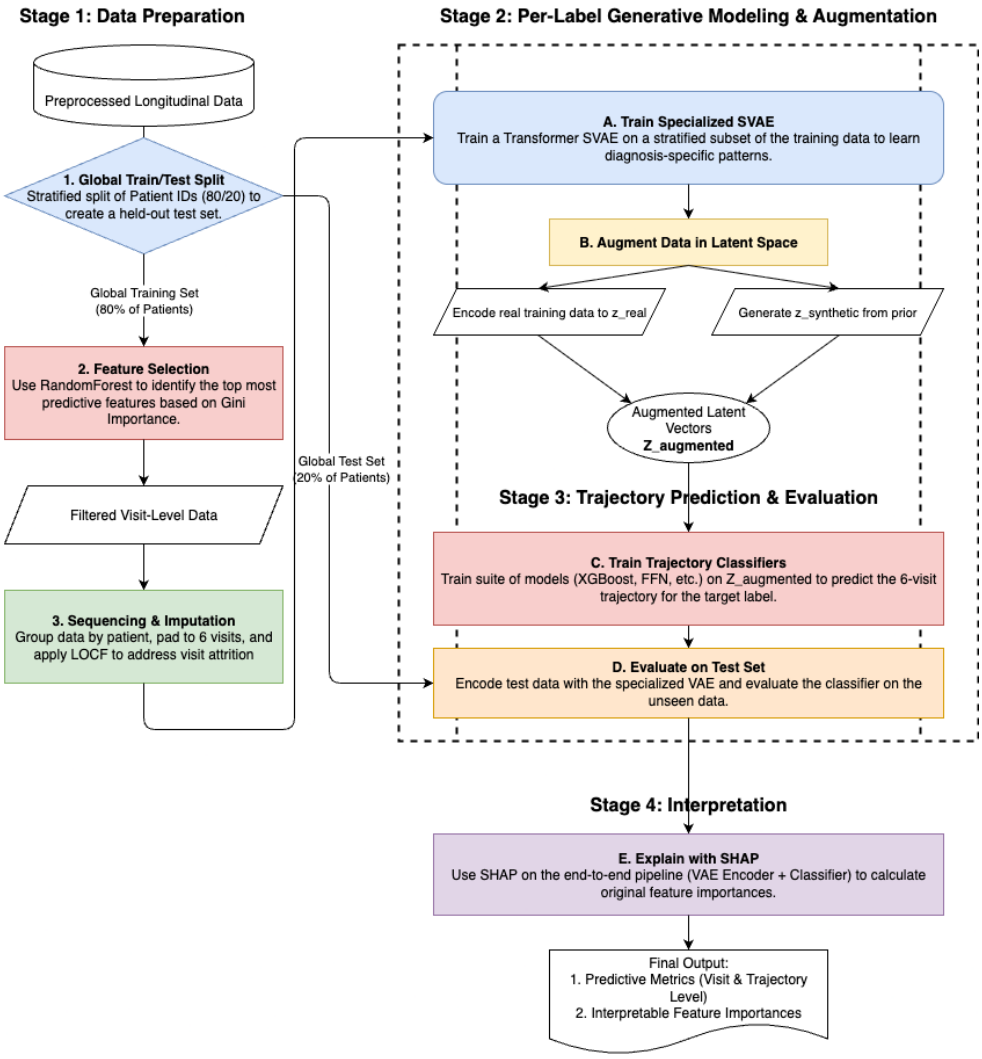

Figure 2: The Four-Stage Research Methodology. Stage 1 involves preparing raw data through feature selection and sequencing. In Stage 2, a specialized SVAE is trained for a target diagnosis and subsequently used to generate an augmented set of latent vectors. In Stage 3, a suite of classifiers is trained and evaluated on this augmented data. Finally, Stage 4 provides model interpretation via SHAP.

B.2 Performance-Based Feature Selection

Figures 3 and 4 show the results of the performance-based feature selection process. Cross-validated ROC-AUC is plotted against the number of features (ranked by Gini importance), empirically identifying the optimal feature set size where performance peaks before declining due to noise.

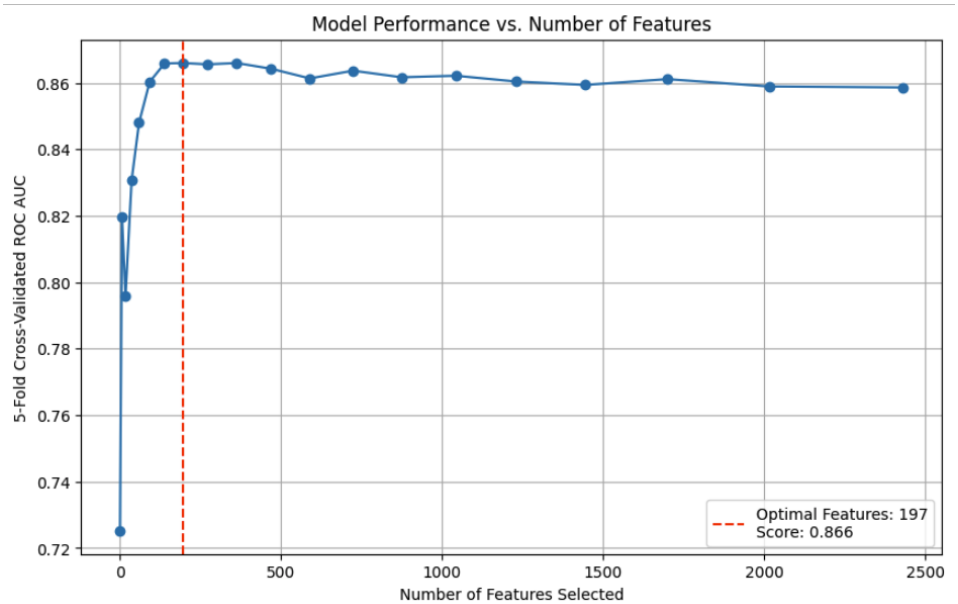

Figure 3: Optimal feature set analysis for the combined Clinical+SDoH dataset. Performance peaks at 197 features, which captures 35% of the cumulative Gini importance and achieves a cross-validated ROC-AUC of 0.866.

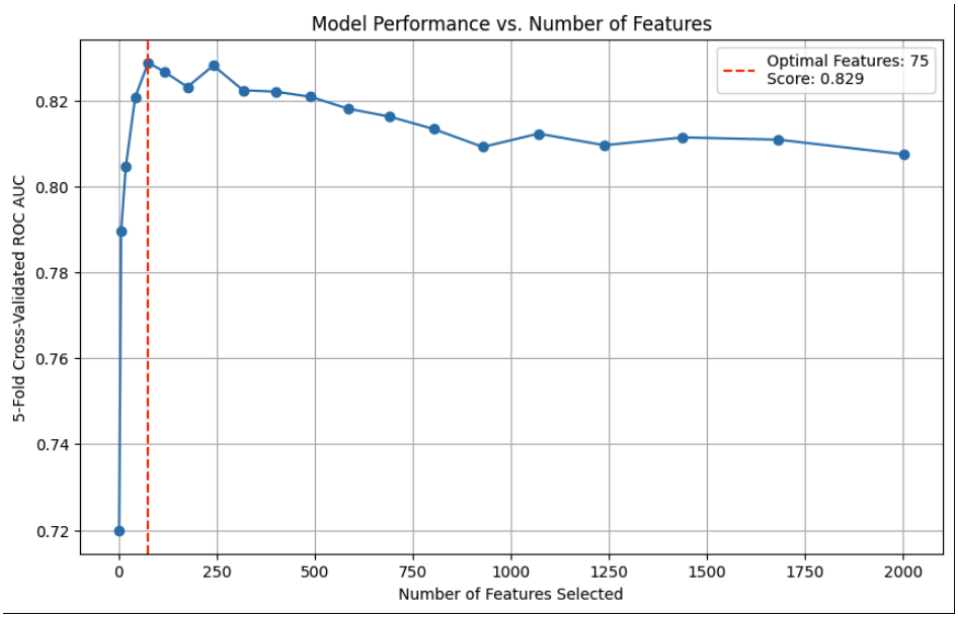

Figure 4: Optimal feature set analysis for the SDoH-only dataset. Performance peaks with just 75 features (25% of cumulative Gini importance), reaching a cross-validated ROC-AUC of 0.829.

B.3 Outcome-Specific SHAP Explanations

The following figures 5, 6, 7 and 8 present the outcome-specific SHAP analysis for the best-performing XGBoost model. By stratifying explanations, we can compare the feature contributions for correct predictions (TPs, TNs) versus incorrect predictions (FPs, FNs), revealing the drivers of model success and failure.

C Reproducibility

Training on an 10-cores 16GB RAM Mac M4 with parallelization and/or GPU cores enabled were possible takes around 2 hrs to run the preprocessing pipeline, 4 hrs to run the full modeling pipeline for each scenario (SDoH only, SDoH + Clinical, with and without free-bits, with and without Conditionality) and the SHAP values generation code takes around 6 hrs to run, also for each scenario. The code will be made public and open source on GitHub for the camera-ready version.

The dataset used can be requested, upon completion of the Human Subjects - Social Behavioral Researchers course, and obtaining the IRB approval from one of the members of the TX-CTRN.

The full codebase used on this research can be accessed on the link https://github.com/matglima/S2SVAE_SynthData

D Limitations

While this study presents a robust framework for psychiatric trajectory modeling, we acknowledge several limitations that provide important directions for future work.

Generalizability. The training and evaluation data were sourced from a single, comprehensive research network (TX-CTRN). Although this ensured high data quality and consistency, the learned predictive patterns and feature importances may not generalize perfectly to different demographic populations or healthcare systems with varying data collection protocols. Future work should aim to validate this framework on more diverse, multi-site datasets to establish its broader applicability.

Imputation Strategy. To handle patient attrition and missed visits, we employed a Last Observation Carried Forward (LOCF) strategy. While practical and effective, LOCF assumes a static state between observed visits, which may not fully capture the dynamic nature of symptom progression. Exploring more sophisticated imputation techniques, such as those based on Gaussian processes or learned by the model itself, could further enhance the fidelity of the patient trajectories.

Privacy Evaluation. Our work focused on demonstrating the fidelity and utility of the synthetic data, evidenced by the strong performance of the downstream classifiers. However, we did not conduct a formal, quantitative evaluation of the privacy-preserving properties of the generated data. For real-world clinical deployment, ensuring that synthetic data does not leak sensitive patient information is critical. Future iterations should incorporate formal privacy assessments, such as calculating the Distance to Closest Record (DCR) or testing resilience against Membership Inference Attacks (MIAs), to provide a complete Fidelity-Utility-Privacy analysis.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Conclusion

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Methodology and Conclusion

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See Results and Discussions

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: Code will be released after double-blind review

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Code will be released after double-blind review, but data access depends on IRB approval

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Methodology

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Methodology and Discussions

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Reproducibility on appendix

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The Code of Ethics was respected and followed for this research.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Introduction and Conclusion

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No pretrained models will be released, only the code

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The license will be added on the code repo

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The necessary code will be released on the Github repository of this research.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: The guidelines are not included on the paper, but are included on the TX-CTRN guidelines, and must be reviewed prior to data access.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The TX-CTRN study is supervised and was approved by UT Austin's IRB.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs are not core methods of this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

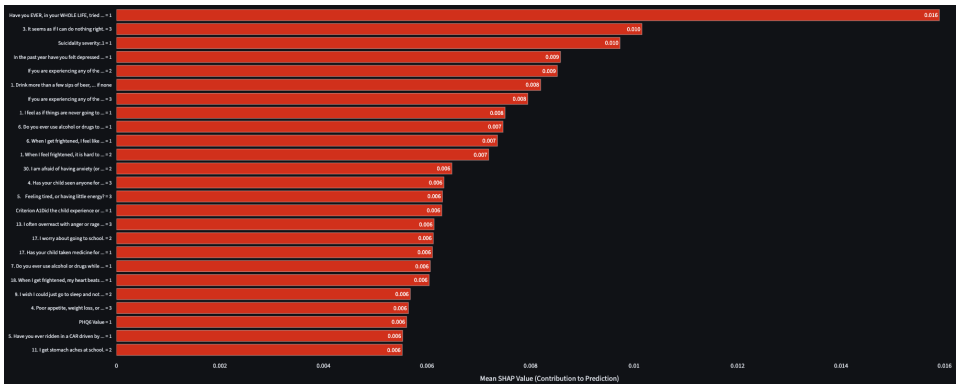

Figure 5: Top 20 features for **True Positive (TP)** predictions.

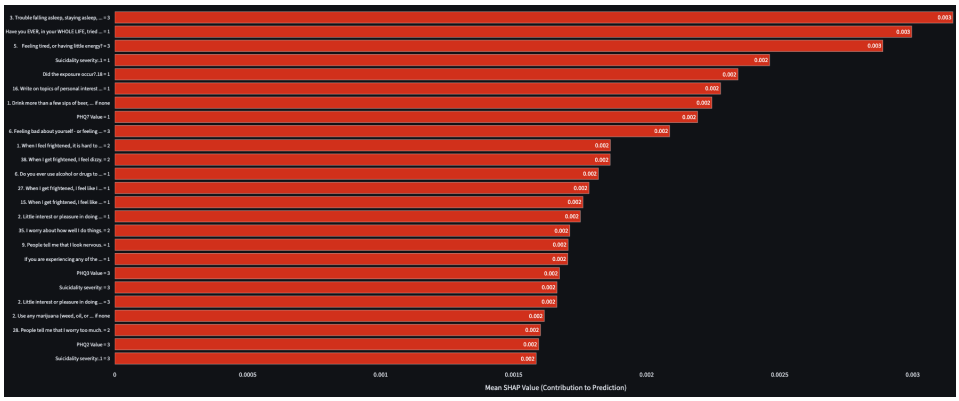

Figure 6: Top 20 features for **False Negative (FN)** predictions.

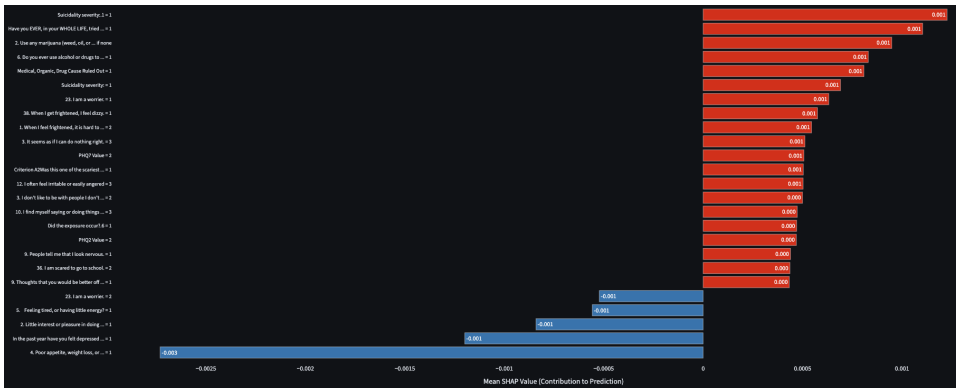

Figure 7: Top 20 features for **True Negative (TN)** predictions.

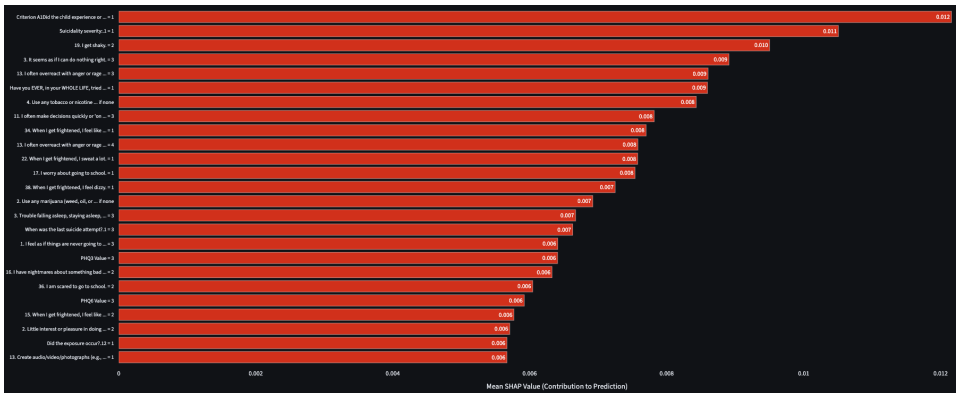

Figure 8: Top 20 features for **False Positive (FP)** predictions.