ISSN: 2583-9292 (ONLINE)
DOI: 10.21917/ijdsml.2025.0165

HEALTHODE: NEURAL DISCRETIZED DYNAMICS FOR ROBUST FORECASTING OF PATIENT WELLNESS

Shrutibahen Patel
Master of Public Administration, School of Public and Global Affairs, Fairleigh Dickinson University, United States of America

Abstract

Electronic health records (EHRs) are inherently irregularly sampled, posing significant challenges for conventional time-series models. In this work, we introduce HealthODE—a novel framework that leverages neural discretized ordinary differential equations (ODEs) to learn robust representations from irregular health data. By integrating a decay-gated attention mechanism and rotary positional encoding, HealthODE adaptively filters irrelevant historical data while accurately capturing continuous dynamics. Our approach supports both interpolations within observed intervals and extrapolation beyond them, enabling zero-shot forecasting for a range of clinical tasks such as diagnostic prediction, drug usage estimation, and phenotype classification. Empirical evaluations demonstrate that HealthODE not only improves forecasting accuracy but also provides interpretable insights into patient risk trajectories, making it a promising tool for advanced healthcare analytics.

Keywords:

HealthODE, Historical data, Zero-Shot Forecasting, Clinical Task

1. INTRODUCTION

Time-series portrayal learning assumes a vital part in different spaces, as it works with the extraction of generalizable fleeting examples from huge-scope, unlabeled information, which can then be adjusted for assorted errands. Be that as it may, a significant test emerges while managing unpredictably tested time series, in which perceptions happen at lopsided spans. This abnormality presents difficulties for old-style time-series models that are confined to normal examination. This issue is especially critical in the medical services space since longitudinal electronic wellbeing records (EHRs) are refreshed irregularly during short-term visits or ongoing stays. In addition, individual clinical chronicles frequently length a restricted period because of an absence of verifiable digitization, deficient protection inclusion, and, what's more, divided medical services frameworks. These difficulties make it hard for time series models to catch the hidden direction elements [1].

Tending to these difficulties requires the improvement of novel portrayal learning methods that can separate generalizable transient examples from unpredictably tested information through next token forecast pre-preparing. The pre-prepared model is then applied to estimate directions in light of the learned adaptable. examples, even when patient information is just noticed to some extent. Ongoing advances in displaying sporadically examined time series have been accomplished through profound learning structures. Be that as it may, these models miss the mark in pre-preparing generalizable representations [2]. While time-series Transformer models stand out enough to be noticed, they are fundamentally intended for continuous information and neglect to represent sporadic spans between perceptions [3]. To deal with both customary and unpredictable time series, Timely GPT consolidates relative position implanting to catch positional data in fluctuating time holes. BiTimelyGPT broadens this by pre-preparing bidirectional portrayals for discriminative errands. Regardless of these enhancements, the two models depend on an information-free rot, which isn't content-mindful and, in this manner, can't completely catch complex worldly conditions in medical care information. The key test stays to foster a viable portrayal-gaining approach that extricates significant examples from unpredictably tested data [4]. In this review, The Direction Generative Pre-Trained Transformer (TrajGPT) is proposed for irregular time-series representation learning. the exploration offers four significant commitments: First, it presents a Particular Repetitive Consideration (SRA) component with an information subordinate rot, empowering the model to adaptively fail to remember immaterial past data in view of settings. Second, by deciphering TrajGPT, as discretized Tributes, catches the consistent elements in unpredictably tested information. This empowers TrajGPT to perform addition and extrapolation in the two headings, permitting a clever time frame explicit derivation for precise estimating. Third, TrajGPT shows solid zero-shot execution across numerous assignments, including direction gauging, drug use forecast, and aggregate arrangement. Finally, TrajGPT offers interpretable wellbeing direction examination, empowering clinicians to adjust the direction of extrapolated sickness movement to hidden patient circumstances.

2. RELATED WORK

2.1 TIME-SERIES TRANSFORMER MODELS

Time-series Transformer models have demonstrated strong performance in modeling temporal dependencies through attention mechanisms. The informer introduces ProbSparse self-attention to extract key information by halving cascading layer input. Autoformer utilizes Autocorrelation to capture series-wise temporal dependencies [5]. FEDformer adopts Fourier-enhanced attention to capture frequency-domain relationships. PatchTST compresses time series into patches and forecasts all timesteps using a linear layer. Despite their effectiveness, these methods fail to account for irregular time intervals. TimelyGPT and BiTimelyGPT address this limitation by encoding irregular time gaps with relative position embedding. However, these models rely on data-independent decay, whereas TrajGPT introduces data-dependent decay to forget irrelevant information based on contexts adaptively. PrimeNet designs a time-sensitive contrastive learning and a masking-and-reconstruction task for irregular time-series representation learning. ContiFormer integrates ODEs into attention and value matrices to model continuous dynamics. However, it demands significantly more computing resources than A standard Transformer with quadratic complexity due to the slow process of solving ODEs [6]. In contrast, TrajGPT models continuous dynamics by pre-training

on irregularly-sampled data with efficient linear training complexity and constant inference complexity.

2.2 ALGORITHMS DESIGNED FOR IRREGULARLYSAMPLED TIME SERIES

Various techniques have been developed to model irregular temporal dependencies through specialized architectures. GRU-D captures temporal dependencies by applying exponential decay to hidden states, [7]. SeFT adopts a set function-based approach, where each observation is modeled individually and then pooled together. RAINDROP captures irregular temporal dependencies by representing data as separate sensor graphs. mTAND employs a multi-time attention mechanism to learn irregular temporal dependencies. In continuous-time approaches, neural ODEs use neural networks to model complex ODEs offer promising interpolation and extrapolation solutions. ODERNN further enhances this by updating RNN hidden states with new observations. HeTVAE addresses sparse input with an uncertainty-aware multi-time attention network and represents variable uncertainty through a heteroscedastic output layer. MGP-TCN combines a multi-task Gaussian Process to manage non-uniform sampling frequencies with temporal convolution network to capture temporal dependencies. However, these Methods lack a representation learning paradigm and often struggle to capture evolving dynamics in partially observed data. In contrast, the TrajGPT can be interpreted as discretized ODEs, allowing it to learn continuous dynamics via largescale pre-training. Moreover, TrajGPT utilizes interpolation and extrapolation techniques from the neural ODE family to predict accurate trajectories.

In addition to existing approaches in irregular time-series modeling, several works from related fields provide complementary insights that can inform and inspire future enhancements to HealthODE.

Cherukuri [8] presents a novel approach for denoising binary images using simulated annealing (SA). Although this work focuses on image segmentation, its use of global optimization to minimize a non-convex energy function parallels the challenges encountered in learning continuous dynamics from sparse, irregular data. The optimization strategies introduced in this paper can offer valuable perspectives on efficiently navigating complex solution spaces in highdimensional medical data.

Similarly, Awasthi [9] performs a comparative analysis of advanced reinforcement learning algorithms—specifically DQN, DDQN, DDPG, and PPO—applied to the LunarLanderv2 control task. By examining the trade-offs between sample efficiency, stability, and computational overhead, this study highlights key design principles for sequential decisionmaking in dynamic environments. These insights are particularly relevant for forecasting health trajectories, where managing temporal dependencies and balancing computational complexity are critical.

Furthermore, Malhotra [10] introduces the Self-Organizing Interaction Spaces (SOIS) framework, which is designed to support pervasive applications in mobile and distributed settings. The adaptive and decentralized architecture proposed in this work can be leveraged to enhance healthcare analytics, especially in scenarios involving real-time data from distributed mobile devices. This approach underlines the potential benefits of integrating self-organizing principles into healthcare systems to improve scalability and responsiveness.

Collectively, these interdisciplinary contributions underscore the potential for cross-domain innovations in representation learning and dynamic modeling, offering promising avenues for future research in healthcare analytics.

3. METHODOLOGY: A CLEARER VIEW OF HEALTHODE

In HealthODE, an irregular time series is represented as $\mathbf{x} = \{(x_1,t_1),(x_2,t_2),\ldots,(x_N,t_N)\}$, where each sample (x_n,t_n) corresponds to a clinical observation recorded at time t_n. This section details the architecture and key components that enable HealthODE to capture complex temporal dynamics in irregularly-sampled healthcare data.

3.1 INPUT EMBEDDING WITH TEMPORAL ENCODING

Each clinical observation x_n is first projected into a highdimensional token embedding space. To account for the nonuniform time intervals between observations, HealthODE employs *Rotary Positional Encoding* (RoPE). This mechanism explicitly encodes the relative temporal distances between tokens by rotating the token embeddings according to their associated timestamps. Such encoding ensures that the model differentiates between events that occur in rapid succession and those separated by longer durations, preserving essential temporal context.

3.2 SEQUENTIAL RECURRENT ATTENTION (SRA) LAYERS

The token embeddings are then processed through multiple SRA layers, which are at the core of the HealthODE architecture. Each SRA layer integrates a data-dependent decay mechanism with standard attention operations to dynamically focus on relevant historical information.

- *Decay Gating::* A learnable decay vector γ_n is introduced to modulate the influence of past observations. This mechanism allows the model to:
- *Preserve long-term trends*: For chronic conditions, the decay is slowed, ensuring that older but relevant information is retained.
- *Emphasize recent events*: For acute conditions, a rapid decay prioritizes more recent observations.
- *Recurrent vs. Parallel Form::* Although the update above is defined recursively, it is mathematically equivalent to a parallel formulation. This equivalence allows for efficient computation while still handling irregular intervals robustly.

3.2.1 Discretized Neural ODE Interpretation:

HealthODE reinterprets the SRA layers as a discretized version of a continuous-time Ordinary Differential Equation (ODE). This perspective enables the model to perform:

1. **Interpolation**: Estimating values within the observed time window.
2. **Extrapolation**: Forecasting values beyond the observed window.

Thus, bridging traditional ODE modeling with modern attention mechanisms.

3.3 INFERENCE STRATEGIES

HealthODE employs two distinct inference strategies:

- **Auto-regressive Inference**: This mode generates sequential predictions at fixed intervals, where each prediction is conditioned on the previous outputs. It follows the traditional transformer-style approach.
- **Time-specific Inference**: Predictions are made directly at arbitrary time points with constant computational complexity. Leveraging the continuous dynamics of $S(t)$, this strategy enhances forecasting accuracy, especially for irregularly spaced data.

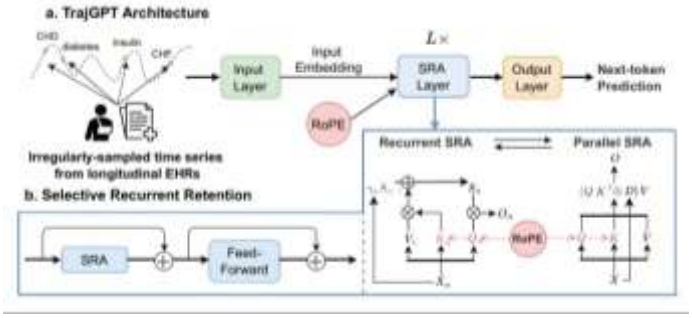

Fig.1. Input Embedding and Temporal Encoding: This figure illustrates how HealthODE embeds irregular time-series data into token space. It demonstrates the application of RoPE to input tokens, emphasizing the encoding of relative temporal distances that differentiate closely spaced events from those separated by longer intervals

3.4 TRAJGPT AS DISCRETIZED ODES

In this section, Theoretical connections are established between the proposed SRA module and ODEs. The recurrent form of SRA is a discretization of continuous time ODE using the zero-order hold (ZOH) rule. Given a first-order ODE, the recurrent SRA can be derived using a ZOH discretization with a discrete step size.

$$C = Q_t,\ \Lambda_t = \mathrm{diag}(1, \gamma_t). \quad (1)$$

This ODE naturally models the continuous dynamics underlying irregularly sampled data, with corresponding to the varying time intervals between observations. Since the parameters (A, B, C) depend on the t^{th} observation $X(t)$; this continuous-time model becomes a neural ODE with a differentiable neural network f and data-dependent parameters $t = (A, B, C)$. Consequently, a single-head SRA serves as a discretized ODE with datadependent Parameters (i.e., neural ODE). TrajGPT with multihead SRA operates as discretized ODEs, where each head corresponds to its own ODE and captures distinct dynamics.

As illustrated in Fig.2(a), TrajGPT functions as discretized ODEs, enabling both interpolation and extrapolation of irregular timeseries data. By capturing the underlying continuous dynamics, TrajGPT handles irregular input through discretization with varying step sizes. For interpolation, the dynamics within the observed time frame evolve using a unit discretization step size. For extrapolation, the dynamics evolve forward or backward in time beyond the observed time frame. Additionally, TrajGPT estimates disease risk trajectories by computing token probabilities at specific time steps and changing the dynamics through interpolation and extrapolation.

At inference time, two strategies for forecasting irregularly sampled time series are explored: auto-regressive and time-specific inference (Fig.2(b)). Auto-regressive inference, commonly used by standard Transformer models, makes sequential predictions at equal intervals and selects the target time steps accordingly.

Since TrajGPT functions as discretized ODEs, a novel time-specific inference is introduced to predict arbitrary time steps. To forecast a target time point (x_n, t_n), TrajGPT utilizes both the target timestep t_n and the last observation (x_n, t_n) to predict the corresponding observation x_n. It calculates the target output representation $O_n = Q_n S_n$, taking into account the discrete step size $t_{n,n} = t_n$ and the updated state $S_n = D_{tm,n}\, S_n + K_n V_n$.

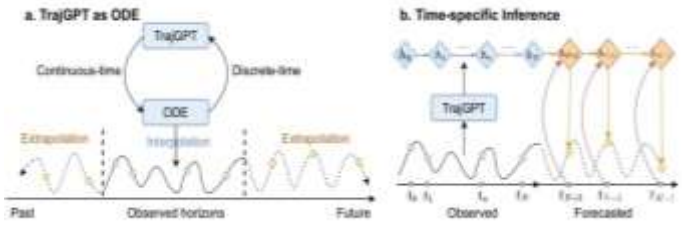

Fig.2. Interpolation and Extrapolation via Discretized Neural ODE: This figure depicts how the discretized neural ODE model facilitates both interpolation within the observed timeframe and extrapolation beyond it. The evolution of the state vector $S(t)$ under varying time steps is highlighted, showcasing the model's capability to forecast future or past states based on learned dynamics

3.5 COMPUTATIONAL COMPLEXITY

TrajGPT, with its efficient SRA mechanics,m achieves linear training complexity of O(N) and constant inference complexity of O(1) with respect to sequence length N. In contrast, standard Transformer models incur quadratic training complexity of $O(N^2)$ and linear inference complexity of O(N) (Katharopoulos et al., 2020). This computational bottleneck arises from the vanilla self-attention mechanism, where Attention(X) = Softmax(QK^T)V, resulting in a training complexity of $O(N^2d)$. When dealing with long sequences, the quadratic term $O(N^2)$ becomes a bottleneck for standard Transformer models.

As a variant of linear attention, the SRA mechanism in TrajGPT overcomes this quadratic bottleneck of the taTransformerformer, achieving linear training complexity for long sequences. By recursively updating over N tokens, the total complexity becomes $O(Nd^2)$. For inference, TrajGPT proposes auto-regressive and time-specific methods.

The auto-regressive inference sequentially generates sequences with equally spaced time intervals like the GPT model, incurring linear complexity of O(N). In contrast, timespecific inference directly predicts the target time point with a constant complexity of O(1). Thus, TrajGPT achieves O(N) training complexity and O(1) inference complexity, making it computationally efficient for long sequences.

4. EXPERIMENTAL DESIGN

4.1 DATASET AND PRE-PROCESSING

Population Health Record (PopHR) data set has monstrous measures of longitudinal case information from the commonplace government well-being backup plan in Quebec, Canada, on well-being administration use. Altogether, there are around 1.3 million members in the PopHR data set, addressing a haphazardly examined 25% of the populace in the metropolitan area of Montreal is somewhere in the range of 1998 and 2014.

Associate participation is kept up progressively by eliminating perished occupants and effectively selecting babies and workers. Sporadically inspected time series were extracted from the PopHR dataset. Specifically, ICD-9 diagnostic codes were converted into integer-level aggregate codes (PheCodes) using the PheWAS list. A total of 194 novel PheCodes were selected, each with more than 50,000 events. Patients with fewer than 50 PheCode records were excluded, resulting in a final dataset of 489,000 patients, with an average of 112 records per person. The dataset was then divided into preparing (80%), approval (10%), and testing (10%) sets. The eICU Cooperative Exploration Data set is a multi-focus emergency unit information base containing north of 200,000 affirmations from ICUs observed by eICU programs in the US. It offers de-recognized EHR information, incorporating patient socioeconomics, judgments, medicines, and medications. To extricate sporadically tested time series, ICD-9 codes were converted to 288 numeric-level PheCodes. Drugs with similar characteristics but varying names and dosages were grouped, resulting in 228 novel medications. Representation learning was performed at a 15-minute interval for clinical events (diagnosis and medication). This brought about a last dataset of 139,367 patients, with a normal of 19 medications, with 3 ICD codes for each tolerance.

4.2 POPHR EXPERIMENT DESIGN

Forecast irregular diagnostic codes The long-term forecasting task was evaluated using a look-up window of 50-time points (e.g., diagnosis codes) to predict the remaining codes in the forecasting windows. Model performance was measured using the top-K recall with K = (5, 10, 15). This metric mimics the behavior of doctors conducting differential diagnoses, where they list the most probable diagnoses based on a patient Choi et al. (2016) [11]. Since next-token prediction is inherently forecasting, TrajGPT enables zero-shot forecasting without requiring fine-tuning. Drug usage prediction In this application, The task involved predicting whether each diabetic patient started insulin treatment within 6 months of their initial diabetes diagnosis. Following the pre-processing from previous work (Song et al., 2021), 78,712 diabetic patients with PheCode 250 were extracted, of which 11,433 patients were labeled as positive. Due to class imbalance, the area under the precision-recall curve (AUPRC) was used as the evaluation metric. To avoid information leakage, sequence representations were truncated at the first diabetes record. To assess generalizability, zero-shot classification, few-shot classification with 5 samples, and fine-tuning on the full dataset were performed.

For phenotype classification, the PopHR database provides rule-based labels for congestive heart failure (CHF), with 3.2% of the total population labeled as positive. Given the class imbalance, the AUPRC metric was used to evaluate performance on the rare positive class. To assess the generalizability of the pre-trained TrajGPT, zero-shot classification, few-shot classification with five samples, and fine-tuning on the entire dataset were also conducted.

4.3 EICU EXPERIMENT DESIGN

Forecast irregular diagnoses and drugs. The forecasting task was conducted using a 10-time point look-up window to predict the remaining codes in the forecasting windows. Forecasting performance was assessed using the top-K recall with K = (10, 20). To detect sepsis early, a 72-hour observation period following ICU admission was defined. Patients without sepsis during the first 8 hours were identified, and sepsis onset was predicted in the remaining windows. This task was performed using both zero-shot learning and fine-tuning on the full dataset.

5. BASELINES

For the PopHR dataset, the model was compared against several time-series transformer baselines. including TimelyGPT, BiTimelyGPT, Informer, Fedformer, AutoFormer, PatchTST, TimesNet, ContiFormer, PrimeNet [12], and Mamba. BiTimelyGPT and PatchTST are encoder-only models that require fine-tuning for forecasting tasks, while other Transformer models with decoders can forecast without additional fine-tuning. Models designed for irregularly sampled time series were also evaluated, including mTAND, GRU-D, RAINDROP, SeFT, ODE-RNN, HeTVAE, and MGP-TCN. For the eICU dataset, TrajGPT was compared. against efficient models from Section 5.2, including TimelyGPT, PatchTST, TimesNet, ContiFormer, PrimeNet, Mamba-2, MTand, and SeFT. Since these models do not have a pre-training method, they were trained from scratch on the training set. Previous works were followed to set Transformer parameters to approximately 7.5 million (Table.5).

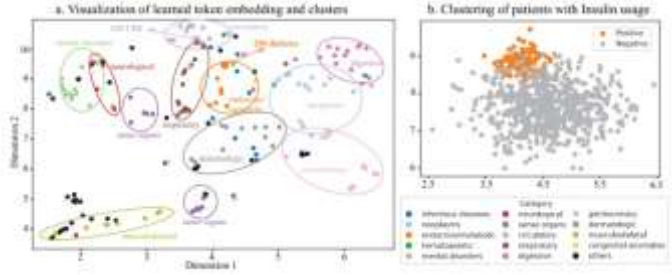

Fig.3. (a) Visualization of token embeddings across 15 disease categories, where token nodes are colored and clustered by categories

Transformer Pre-training paradigm with a cross-entropy loss, TrajGPT employs a next-token prediction task to pre-train generalizable temporal representations from unlabeled data. Given a sequence with a [SOS] token, TrajGPT predicts subsequent tokens by shifting the sequence to the right. The output representation of each token is fed into a linear layer for next-token prediction. For other models without an established pre-training paradigm, A masking-based method was employed by randomly masking 40% timesteps with zeros. All Transformer

models underwent 20 epochs of pre-training with crossentropy loss. When fine-tuning was applicable, five epochs of end-to-end fine-tuning were performed on the entire model.

6. RESULTS

6.1 QUALITATIVE ANALYSIS OF EMBEDDINGS

In this section, A qualitative analysis of the token embeddings and sequence representations learned by TrajGPT on the PopHR database was provided (Fig.3). Uniform Manifold Approximation and Projection (UMAP) was applied to visualize the global token embedding, with nodes colored and clustered by disease categories. The results reveal 12 clearly separated clusters. Some nodes are projected into other categories but still reflect meaningful clinical relationships; for instance, the mental disorders cluster (in green color) includes a black dot representing adverse drug events and drug allergies, implying a high risk of opioid usage among the psychiatric group. Related disease categories with clinical relevance tend to cluster near each other. For example, mental disorders are closely clustered with neurological diseases, and circulatory diseases are adjacent to endocrine/metabolic diseases. The projected head-specific decay vectors w_h are visualized using UMAP techniques. It shows that the eight decay vectors are projected into distinct 2-D vectors, indicating that they capture different patterns. In Fig.3(b), The sequence representations are visualized to demonstrate the ability to perform zero-shot classification of initial insulin usage among diabetic patients. To prevent information leakage, the sequence representations were truncated at the first diabetes record. These sequence representations were projected onto the same scale as the token embeddings in Fig.3(a), allowing for direct comparison with the disease clusters. Patients taking future insulin treatment have embeddings closely aligned with the endocrine/metabolic cluster, indicating a strong association with diabetes-related conditions. In contrast, non-insulin patients are dispersed across various clusters, suggesting less severe diabetes histories. The clear separation between these groups highlights TrajGPT's ability to perform zero-shot classification, showcasing the generalizability of its learned representations.

6.2 QUANTITATIVE RESULTS ON POPHR DATASET

TrajGPT with time-specific inference achieves the highest recall at K = 10 and K = 15, with scores of 71.7% and 84.1%, respectively (Table.1). At K = 5, TrajGPT achieves the secondhighest recall with 57.4%. Notably, time-specific inference outperforms the auto-regressive inference approach, demonstrating its effectiveness in forecasting based on the learned continuous dynamics. These results highlight the strength in pre-training underlying dynamics from sparse and irregular time-series data, facilitating accurate trajectory forecasting over irregular time intervals. The distributions of top 10 recall across three forecast windows are then examined, comparing the two inference methods of TrajGPT as well as TimelyGPT, PatchTST, and mTAND (Fig.6). TrajGPT's time-specific inference consistently outperforms auto-regressive inference as the forecasting window increases, as it accounts for evolving states and query time-steps over irregular intervals. As expected, all models experience a performance decline as the forecast window increases, reflecting the increased uncertainty in long-term trajectory prediction. Despite this, TrajGPT achieves superior and more sTable.performance within the first 100 steps. In comparison, PatchTST shows a drastic decline as the window size increases, reflecting its difficulty with extrapolation. Therefore, TrajGPT excels in forecasting health trajectories through its time-specific inference. Two classification tasks—insulin usage prediction is evaluated, and CHF phenotype classification—under three settings: zeroshot learning, few-shot learning with S = 5 samples, and fine-tuning on the entire dataset. Notably, non-transformer models designed for irregularly sampled time series (i.e., the last five methods in Table. 1) were trained from scratch. The results are summarized in Table. 1. For classification tasks, TrajGPT achieves the highest zero-shot results, with 67.2% for insulin and 72.8% for CHF. This success can be attributed to TrajGPT's ability to learn distinct clusters of sequence representations, as discussed in Section 5.1. For 5shot classification, TrajGPT achieves the second-best performance in both tasks. For fine-tuning, it obtains the second-best performance of 83.9% in insulin prediction, only 0.3% behind the best-performing BiTimelyGPT. A comparison was also made between TrajGPT and algorithms specifically designed for irregularly sampled time series. These methods generally perform worse in insulin usage prediction, likely due to their difficulty in capturing meaningful temporal dependencies from truncated sequences. However, stand outperforms all models in the CHF task, achieving the best result at 85.4%.

6.3 TRAJECTORY ANALYSIS

This analysis aimed to demonstrate effectiveness in trajectory modeling and provide insights into its classification performance. Case studies were conducted on two patients: one diagnosed with diabetes and the other with CHF. The observed and predicted disease trajectories for both patients were visualized, along with the estimated risk trajectories over their lifetimes. As discussed in Section 3.2, risks were interpolated within the observed timeframe and extrapolated beyond it in both directions, with risk computed as the token probability at each timestep. Risk growth was calculated by comparing each timestep to the previous one, identifying ages with high-risk growth, as well as the associated phenotypes. By comparing disease and risk trajectories, phenotype progression, disease comorbidity, and long-term risk development were evaluated. In Fig.4(a), TrajGPT with time-specific inference achieves a top-10 recall of 90.1% for this diabetic patient. TrajGPT accurately predicts most diseases in the endocrine/metabolic and circulatory systems. Although this patient has no prior diabetes diagnosis in the observed data, TrajGPT successfully forecasts diabetes onset by identifying related metabolic and circulatory symptoms. The Fig.4(b) illustrates the predicted risk trajectory for this patient, indicating a gradual increase in diabetes risk with age. Specific phenotypes that contribute to the noticeable high-risk growth are highlighted between ages 59 and 62, including chronic IHD, hypothyroidism, obesity, and arrhythmia [13]. These conditions are common comorbidities of diabetes, substantially elevating the likelihood of diabetes onset over time. In Fig.4(c), The disease trajectory of a CHF patient was visualized, for whom TrajGPT produced a top-10 recall of 84.7%. TrajGPT accurately predicted a broad range of circulatory, respiratory, and endocrine/metabolic diseases. Despite the absence of prior CHF diagnosis, TrajGPT

successfully predicts the onset of CHF based on a series of related circulatory conditions Correale et al. (2020). In Fig.4.d, the expected risk trajectory reveals two spikes in risk growth at ages 65 and 74, corresponding to successive occurrences of circulatory diseases. This analysis demonstrates TrTrajGPT' ability to forecast unseen phenotypes based on disease comorbidity and the risk with age. As a result, TrTrajGPT' ability to model health trajectories and capture disease progression enhances its classification performance. The ability to forecast diagnostic codes highlights the potential of Transformer models for health trajectory analysis. These codes can serve a broad range of administrative purposes, such as estimating the diagnostic related group (DRG) for inpatients to improve the efficiency and quality of inpatient care. They also hold significant potential for informing clinical care, including directing the need for preventive care and identifying potential complications.

6.4 QUANTITATIVE RESULTS ON EICU DATASET

For the eICU datasets, TrajGPT was evaluated on irregular clinical event forecasting (diagnoses and drugs) and early detection of sepsis, with the results summarized in Table.2. Note that the recall values for the joint prediction of diagnoses and drugs are lower due to the larger hypothesis space for this task. Despite the increased complexity compared to predicting diagnoses alone, TrajGPT with time-specific inference achieved superior performance over baseline models, resulting in a top-10 recall of 57.8% and a top-20 recall of 69.3%. This superior performance can be attributed to the effectiveness of time-specific inference, which improves top-10 and top-20 recall rates by 3.7% and metrics are reported as average (standard error) from a bootstrap evaluation of variance. the bold and underline indicate the best and second best results, respectively. s indicates the number of few-shot examples.

The representation learning methods designed specifically for irregularly sampled time series demonstrated better overall performance. Additionally, ODE-RNN achieves the second best performance with a top-20 recall of 67.8%. These findings highlight that both Time-specific inference and ODE-RNN leverage the strengths of modeling underlying dynamics to enhance forecasting accuracy. For the sepsis prediction task, TrajGPT outperforms all baselines in the zero-shot setting, achieving an AUPRC of 45.1%. While MTand performs best when trained from scratch; its reliance on a bespoke shallow model targeting a single outcome limits its scalability and applicability in clinical settings. In summary, TrajGPT leverages pre-trained generalizable patterns to enable zero-shot learning, effectively detecting early sepsis without additional training.

Table.1. Quantitative results on the diagnosis forecasting, insulin usage, and CHF classification performance on the POPHR dataset

Methods / Tasks (%)		Forecasting			Diabetes-Insulin			CHF	
(lr)2-4 (lr)5-7 (lr)8-10	K = 5	10	15	S = 0	5	all	S = 0	5	all
TrajGPT (Time-specific)	57.4 (3.2)	71.7 (2.6)	84.1 (2.4)	67.2 (3.1)	70.2 (3.0)	75.5 (2.6)	72.8 (2.4)	75.9 (2.1)	83.9 (2.0)
TrajGPT (Auto-regressive)	53.3 (3.9)	65.5 (3.4)	77.2 (2.7)	—	—	—	—	—	—
TimelyGPT	58.2 (3.7)	70.3 (3.1)	82.1 (2.5)	58.2 (2.8)	64.4 (2.5)	70.7 (2.6)	66.9 (2.3)	71.0 (2.2)	81.2 (2.0)
BiTimelyGPT	48.2 (3.3)	63.3 (3.2)	70.5 (2.8)	65.3 (3.1)	70.8 (2.9)	75.8 (3.0)	70.4 (2.4)	74.5 (2.3)	83.8 (2.1)
Informer	46.4 (2.9)	60.1 (2.8)	71.2 (2.6)	62.1 (4.6)	66.2 (4.5)	71.5 (3.8)	62.9 (4.2)	67.4 (3.9)	80.8 (3.5)
Autoformer	42.9 (2.9)	57.4 (2.7)	68.6 (2.4)	63.5 (3.8)	66.8 (3.6)	72.7 (3.4)	65.3 (3.5)	69.6 (3.7)	81.6 (3.2)
Fedformer	43.3 (2.7)	58.3 (2.5)	69.6 (2.4)	64.2 (4.3)	68.4 (4.2)	73.1 (3.8)	68.2 (3.8)	69.8 (3.5)	81.9 (2.9)
PatchTST	48.2 (2.7)	65.5 (2.4)	73.3 (2.2)	66.8 (2.6)	69.7 (2.7)	75.1 (2.4)	72.2 (2.3)	76.3 (1.9)	84.2 (2.1)
TimesNet	46.5 (3.7)	64.3 (3.0)	71.5 (2.5)	64.2 (3.2)	67.9 (2.8)	72.8 (2.9)	67.8 (3.1)	72.5 (3.0)	82.6 (2.8)
ContiFormer	52.8 (3.1)	67.2 (2.8)	76.9 (2.5)	63.5 (3.3)	68.0 (3.1)	75.0 (2.9)	68.4 (2.4)	74.9 (2.2)	83.1 (2.3)
PrimeNet	52.5 (3.2)	69.7 (2.8)	81.8 (2.3)	65.6 (3.0)	69.5 (2.9)	73.8 (2.7)	71.5 (2.7)	75.5 (2.9)	84.0 (2.4)
Mamba-1	46.5 (3.6)	62.4 (3.1)	73.6 (2.6)	61.5 (3.6)	67.4 (3.2)	72.5 (3.0)	65.2 (3.1)	70.1 (2.9)	81.4 (2.4)
Mamba-2	51.4 (3.2)	69.8 (2.9)	80.7 (2.5)	64.6 (3.1)	69.9 (2.8)	74.8 (2.4)	69.6 (2.7)	73.9 (2.8)	83.4 (2.3)
MTand	52.6 (2.8)	70.2 (2.5)	83.7 (1.9)	—	—	74.6 (3.1)	—	—	85.4 (2.5)

GRU-D	54.2 (4.0)	69.5 (3.4)	80.5 (3.1)	—	—	72.1 (3.2)	—	—	79.9 (2.7)
RAINDROP	46.5 (2.9)	67.2 (2.5)	72.2 (2.2)	—	—	70.5 (2.8)	—	—	82.4 (2.4)
SeFT	49.3 (2.6)	68.1 (2.2)	79.4 (1.7)	—	—	71.7 (2.6)	—	—	83.4 (2.3)
ODE-RNN	54.7 (4.2)	70.6 (3.5)	78.6 (2.8)	—	—	73.5 (3.6)	—	—	82.9 (3.0)
HeTVAE	51.1 (3.9)	70.1 (3.4)	83.2 (3.2)	—	—	71.4 (3.6)	—	—	81.6 (3.2)
MGP-TCN	43.5 (3.5)	57.2 (3.1)	69.1 (2.9)	—	—	73.9 (3.6)	—	—	82.4 (3.5)

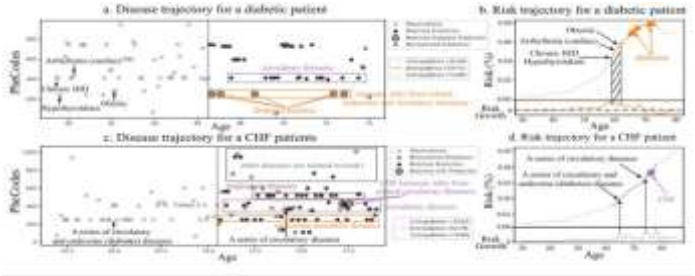

Fig.4. Inferred disease trajectories with look-up and forecast windows. Matched predictions are shown as solid circles, with larger circles for correctly predicted diabetes or CHF

Table.2. Evaluation of TRAJGPT and baselines on the EICU dataset for event forecasting and sepsis prediction. metrics are reported as average (standard error) from a bootstrap evaluation of variance. bold and underlined values indicate the best and second-best results, respectively

Methods/Tasks (%)	**Forecasting**		**Sepsis**	
(lr)2-3 (lr)4-5	K = 10	K = 20	S = 0	All
TrajGPT (Time-specific)	57.8 (2.9)	69.3 (2.1)	45.1 (2.7)	51.3 (2.4)
TrajGPT (Auto regressive)	54.1 (3.2)	64.9 (2.3)	-	-
TimelyGPT	56.9 (3.2)	67.1 (2.4)	42.0 (2.5)	48.5 (2.2)
PatchTST	55.2 (2.7)	66.0 (1.7)	44.5 (2.2)	51.8 (1.8)
TimesNet	52.9 (3.1)	60.3 (2.3)	41.2 (3.1)	47.5 (2.6)
ContiFormer	57.1 (2.2)	66.8 (2.2)	41.7 (2.5)	50.6 (2.8)
PrimeNet	53.4 (2.3)	67.5 (2.0)	44.0 (2.3)	51.2 (1.9)
Mamba-2	55.7 (2.8)	65.2 (2.3)	43.6 (2.8)	49.5 (2.3)
MTand	53.9 (2.4)	67.4 (1.6)	-	52.5 (2.1)
ODE-RNN	55.7 (3.4)	67.8 (2.8)	-	49.2 (2.9)

Table.3. Ablation results of TRAJGPT by selectively removing components and comparing inference methods. performance is evaluated on the forecasting task with the top 10 recall

Model Variants	**Time-specific**	**Auto Regressive**
	Inference	**Inference**
TrajGPT	71.7	65.5
w/o decay gating (i.e., fixed γ)	70.3	64.0
w/o RoPE (i.e., absolute PE)	67.8	63.2
w/o linear attention (i.e., GPT-2)	-	61.2
TrajGPT (without Pre-training)	67.1	?

6.5 ABLATION STUDY

To evaluate the contributions of key components in TrajGPT, Ablation studies were performed by selectively removing elements such as decay gating, RoPE, and the linear attention module. The time-specific inference and auto-regressive inference were compared under different ablation setups. Notably, removing all components results in a vanilla GPT-2, which is limited to performing only auto-regressive inference. The ablation studies were assessed on the forecasting task using the top-10 recall metric. As shown in Table.3, removing the data dependent decay and RoPE results in performance declines of 1.4% and 2.5%, respectively. This highlights the critical role of these modules in handling irregular time intervals by prioritizing recent data while attenuating the influence of distant ones. Replacing time-specific inference with auto-regressive inference leads to performance drops ranging from 4.6% to 6.2%, with the most significant drop in TrajGPT. Furthermore, vanilla GPT-2 with auto-regressive inference produces the lowest performance, falling behind TrajGPT with time-specific inference by 10.5%. Time-specific inference uses varied time intervals for a single inference, reducing both computational steps and error accumulation for better performance.

7. CONCLUSION AND FUTURE WORK

The ongoing worldview in clinical practice depends on custom-tailored shallow models focusing on single results, featuring the requirement for models equipped for anticipating assorted patient results with insignificant or no refinement. Growing such models for medical services needs to represent the unpredicTable.examining of clinical records, as inappropriate displaying can prompt personnel derivations. the exploration proposes an original engineering, TrajGPT, intended for sporadic timeseries portrayal learning and examination. To accomplish this, TrajGPT presents an SRA component with an information subordinate rot, permitting the model to fail to remember unimportant past specifically data in light of settings. Deciphering TrajGPT as discretized Tributes really catches the constant elements of fundamental sporadically tested time series, empowering both introduction and extrapolation. For the determining task, TrajGPT gives a powerful time-explicit

derivation by advancing the elements as indicated by differing periods. TrajGPT exhibits solid zeroshot execution across various assignments, including conclusion anticipating, drug use expectation, and aggregate arrangement. TrajGPT additionally gives interpreTable.direction examination, helping clinicians understand the extrapolated infection movement alongside risk development. Additionally, to approve generalizability, the work focuses on irregularly sampled time series with discrete data (e.g., diagnoses and medications). It plans to extend it to continuous multivariate time series, such as ICU measurements. Future research will also explore representation learning and trajectory analysis on out-of-distribution data.

REFERENCES

[1] A. Amirahmadi, M. Ohlsson and K. Etminani, "Deep Learning Prediction Models based on EHR Trajectories: A Systematic Review", *Journal of Biomedical Informatics*, Vol. 144, pp. 1-7, 2023.

[2] Z. Che, S. Purushotham, K. Cho, D. Sontag and Y. Liu, "Recurrent Neural Networks for Multivariate Time Series with Missing Values", *Scientific Reports*, Vol. 8, pp. 1-12, 2018.

[3] J.R.A. Solares, F.E.D. Raimondi, Y. Zhu, F. Rahimian, D. Canoy, J. Tran, A.C.P. Gomes, A.H. Payberah, M. Zottoli, M. Nazarzadeh, N. Conrad, K. Rahimi and G. Salimi-Khorshidi, "Deep Learning for Electronic Health Records: A Comparative Review of Multiple Deep Neural Architectures", *Journal of Biomedical Informatics*, Vol. 101, pp. 1-15, 2020.

[4] D. Agniel, I. Kohane and G. Weber, "Biases in Electronic Health Record Data Due to Processes within the Healthcare System: Retrospective Observational Study", *BMJ*, Vol. 361, pp. 1-6, 2018.

[5] M. Correale, S. Paolillo, V. Mercurio, G. Limongelli, F. Barilla, G. Ruocco, A. Palazzuoli, D. Scrutinio, R. Lagioia, C. Lombardi, L. Lupi, D. Magri, D. Masarone, G. Pacileo, P. Scicchitano, M.M. Ciccone, G. Parati, C.G. Tocchetti and S. Nodari, "Comorbidities in Chronic Heart Failure: An Update from Italian Society of Cardiology (SIC) Working Group on Heart Failure", *European Journal of Internal Medicine*, Vol. 71, pp. 23-31, 2020.

[6] R.T.Q. Chen, Y. Rubanova, J. Bettencourt and D. Duvenaud, "Neural Ordinary Differential Equations", *Proceedings of International Conference on Neural Information Processing Systems*, pp. 6572-6583, 2018.

[7] Y. Chen, K. Ren, Y. Wang, Y. Fang, W. Sun and D. Li, "Contiformer: Continuous-Time Transformer for Irregular Time Series Modeling", *Proceedings of International Conference on Neural Information Processing Systems*, pp. 1-33, 2024.

[8] M. Cherukuri, "Comparing Image Segmentation Algorithms", Proceedings of International Conference on Data Science and Computer Application, pp. 266-269, 2024.

[9] A. Awasthi, "Evaluating Reinforcement Learning Algorithms for Lunarlander-v2: A Comparative Analysis of DQN, DDQN, DDPG and PPO", Available at https://doi.org/10.21203/rs.3.rs-5939959/v1, Accessed in 2025.

[10] S. Malhotra, "Self-Organizing Interaction Spaces: A Framework for Engineering Pervasive Applications in Mobile and Distributed Environments", Available at https://arxiv.org/abs/2502.01137, Accessed in 2025.

[11] E. Choi, M.T. Bahadori, A. Schuetz, W.F. Stewart and J. Sun, "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks", *Proceedings of International Conference on Machine Learning Research*, Vol. 56, pp. 301-318, 2016.

[12] R.R. Chowdhury, J. Li, X. Zhang, D. Hong, R.K. Gupta and J. Shang, "Primenet: Pre-Training for Irregular Multivariate Time Series", *Proceedings of International Conference on Artificial Intelligence*, pp. 1-9, 2023.

[13] B. Biondi, G.J. Kahaly and R.P. Robertson, "Thyroid Dysfunction and Diabetes Mellitus: Two Closely Associated Disorders", *Endocrine Reviews*, Vol. 40, No. 3, pp. 789-824, 2019.