Improving Medical Predictions by Irregular Multimodal Electronic Health Records Modeling

Xinlu Zhang *[1] Shiyang Li *[1] Zhiyu Chen [1] Xifeng Yan [1] Linda Ruth Petzold [1]

Abstract

Health conditions among patients in intensive care units (ICUs) are monitored via electronic health records (EHRs), composed of numerical time series and lengthy clinical note sequences, both taken at *irregular* time intervals. Dealing with such irregularity in every modality, and integrating irregularity into multimodal representations to improve medical predictions, is a challenging problem. Our method first addresses irregularity in each single modality by (1) modeling irregular time series by dynamically incorporating hand-crafted imputation embeddings into learned interpolation embeddings via a gating mechanism, and (2) casting a series of clinical note representations as multivariate irregular time series and tackling irregularity via a time attention mechanism. We further integrate irregularity in multimodal fusion with an interleaved attention mechanism across temporal steps. To the best of our knowledge, this is the first work to thoroughly model irregularity in multimodalities for improving medical predictions. Our proposed methods for two medical prediction tasks consistently outperforms state-of-the-art (SOTA) baselines in each single modality and multimodal fusion scenarios. Specifically, we observe relative improvements of 6.5%, 3.6%, and 4.3% in F1 for time series, clinical notes, and multimodal fusion, respectively. These results demonstrate the effectiveness of our methods and the importance of considering irregularity in multimodal EHRs. [1].

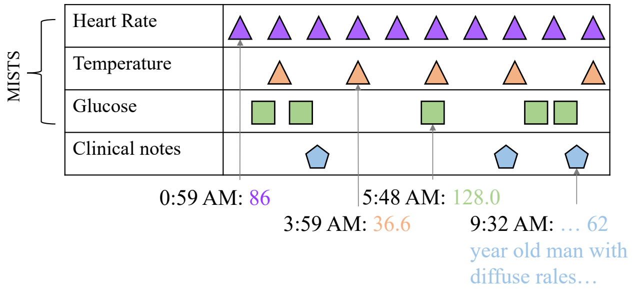

Figure 1: An example of a patient's ICU stay includes MISTS with three features and a series of clinical notes. For MISTS, heart rate and temperature are monitored regularly with different frequencies, and glucose is a laboratory test ordered at irregular time intervals based on doctors' decisions. Clinical notes are free text, collected with much sparser irregular time points than clinical measurements.

1. Introduction

ICUs admit patients with life-threatening conditions, e.g. trauma (Tisherman & Stein, 2018), sepsis (Alberti et al., 2002), and organ failure (Afessa et al., 2007). Care in the first few hours after admission is critical to patient outcomes. This period is also more prone to medical decision errors than later times (Otero-López et al., 2006). Automated tools with effective and real-time predictions can be much beneficial in assisting clinicians in providing appropriate treatments. Recently, the health conditions of patients in ICUs have been recorded in EHRs (Adler-Milstein et al., 2015), bringing the possibility of applying deep neural networks to healthcare (Xiao et al., 2018; Shickel et al., 2017), e.g. mortality prediction (Zhang et al., 2021a) and phenotype classification (Harutyunyan et al., 2019). EHRs contain multivariate irregularly sampled time series (MISTS) and irregular clinical note sequences, as shown in Figure 1. The multimodal structure and complex irregular temporal nature of the data present challenges for prediction. This leads us to formulate two research objectives:

1. *Tackling irregularity in both time series and clinical notes*
2. *Integrating irregularity into multimodal representation learning*

*Equal contribution [1]University of California, Santa Barbara. Correspondence to: Xinlu Zhang <xinluzhang@ucsb.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] Our code is released at https://github.com/XZhang97666/MultimodalMIMIC

To the best of our knowledge, none of the existing works has fully considered irregularity in multimodal representation learning.

We observed three major drawbacks for irregular multimodal EHRs modeling in existing works. 1) *MISTS models perform diversely.* While the numerous MISTS models have been proposed to tackle irregularity (Lipton et al., 2016; Shukla & Marlin, 2019; 2021; Zhang et al., 2021b; Horn et al., 2020; Rubanova et al., 2019), none of the approaches consistently outperforms the others. Even among ***T**emporal **d**iscretization-based **e**mbedding* (TDE) methods, including hand-crafted imputation (Lipton et al., 2016) and learned interpolation (Shukla & Marlin, 2019; 2021), which transform MISTS into regular time representations to interface with deep neural networks for regular time series, there is no clear superior approach. 2) *Irregularity in clinical notes is not well tackled.* Most existing works (Golmaei & Luo, 2021; Mahbub et al., 2022) directly concatenate all clinical notes of each patient but ignore the note-taking time information. Although Zhang et al. (2020) proposes an LSTM variant to model time decay among clinical notes, this approach utilizes only a few trainable parameters, which could be less powerful. 3) *Exiting works ignore irregularity in multimodal fusion.* Deznabi et al. (2021); Yang et al. (2021) have demonstrated the effectiveness of combining time series and clinical notes for medical prediction tasks, however these works are deployed only on multimodal data without considering irregularity. Their fusion strategies may not be able to fully integrate irregular time information into multimodal representations, which can be essential for prediction performance in real-world scenarios.

Our Contributions. To tackle the aforementioned issues, we separately model irregularity in MISTS and irregular clinical notes, and further integrate multimodalities across temporal steps, so as to provide powerful medical predictions based on the complicated irregular time pattern and multimodal structure of EHRs. Specifically, we first show that different TDE methods of tackling MISTS are complementary for medical predictions, by introducing a gating mechanism that incorporates different TDE embeddings specific to each patient. Secondly, we cast note representations and note-taking time as MISTS, and leverage a time attention mechanism (Shukla & Marlin, 2021) to model the irregularity in each dimension of note representations. Finally, we incorporate irregularity into multimodal representations by adopting a fusion method that interleaves self-attentions and cross-attentions (Vaswani et al., 2017) to integrate multimodal knowledge across temporal steps. To the best of our knowledge, this is the first work for a unified system that fully considers irregularity to improve medical predictions, not only in every single modality but also in multimodal fusion scenarios. Our approach demonstrates superior performance compared to baselines in both single modality and multimodal fusion scenarios, with notable relative improvements of 6.5%, 3.6%, and 4.3% in terms of F1 for MISTS, clinical notes, and multimodal fusion, respectively. Our comprehensive ablation study demonstrates that tackling irregularity in every single modality benefits not only their own modality but also multimodal fusion. We also show that modeling long sequential clinical notes further improves medical prediction performance.

2. Related Work

Multivariate irregularly sampled time series (MISTS). MISTS refer to observations of each variable that are acquired at irregular time intervals and can have misaligned observation times across different variables (Zerveas et al., 2021). GRU-D (Che et al., 2018) captures temporal dependencies by decaying the hidden states in gated recurrent units. SeFT (Horn et al., 2020) represents the MISTS to a set of observations based on differentiable set function learning. ODE-RNN (Rubanova et al., 2019) uses latent neural ordinary differential equations (Chen et al., 2018) to specify hidden state dynamics and update RNN hidden states with a new observation. RAINDROP (Zhang et al., 2021b) models MISTS as separate sensor graphs and leverages graph neural networks to learn the dependencies among variables. These approaches model irregular temporal dependencies in MISTS from different perspectives through specialized design. TDE methods are a subset of methods for handling MISTS, converting them to fixed-dimensional feature spaces, and feeding regular time representations into deep neural models for regular time series. Imputation methods (Lipton et al., 2016; Harutyunyan et al., 2019; McDermott et al., 2021) are straightforward TDE methods to discretize MISTS into regular time series with manual missing values imputation, but these ignore the irregularity in the raw data. To fill this gap, Shukla & Marlin (2019) presents interpolation-prediction networks (IP-Nets) to interpolate MISTS at a set of regular reference points via a kernel function with learned parameters. Shukla & Marlin (2021) further presents a time attention mechanism with time embeddings to learn interpolation representations. However, learned interpolation strategies do not always outperform simple imputation methods. This may be due to complicated data sampling patterns (Horn et al., 2020). Inspired by Mixture-of-Experts (MoE) (Shazeer et al., 2017; Jacobs et al., 1991), which maintains a set of experts (neural networks) and seeks a combination of the experts specific to each input via a gating mechanism, we leverage different TDE methods as submodules and integrate hand-crafted imputation embeddings into learned interpolation embeddings to improve medical predictions.

Irregular clinical notes modeling. (Golmaei & Luo, 2021; Mahbub et al., 2022) concatenate each patient's clinical

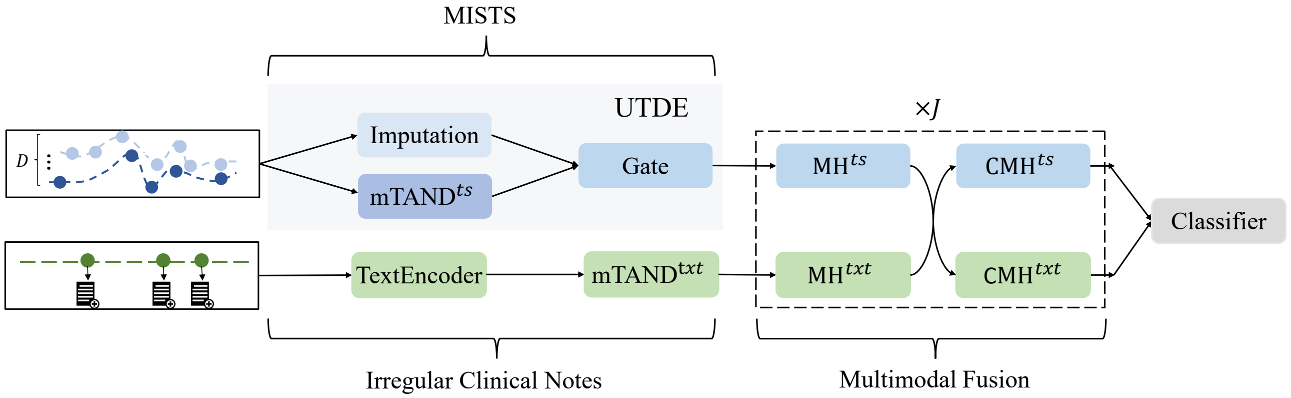

Figure 2: The model architecture, which encodes MISTS and clinical notes separately, and then performs a multimodal fusion. UTDE is a gating mechanism to obtain MISTS representations by dynamically fusing embeddings of imputation and a time attention module, mTAND^{ts}. Irregular clinical notes are encoded by a pretrained language model, $\mathrm{TextEncoder}$, whose outputs are fed into mTAND^{txt} to obtain text interpolation representations. The multimodal fusion strategy contains J identical layers. Each layer interleaves self-attentions (MH) and cross-attentions (CMH) to integrate representations from multimodalities and incorporate irregularity into multimodal representations. A classifier with fully connected layers is used to predict patient outcomes.

notes, divide them into blocks, and then obtain text representations by feeding a series of note blocks into BERT (Devlin et al., 2018) variants (Huang et al., 2019; Gu et al., 2021), ignoring the irregularity in clinical notes. Zhang et al. (2020) further proposes a time-awarded LSTM with trainable decay function to model irregular time information among clinical notes. However, this approach can be less powerful due to limited parameters. To fully model irregularity, we cast clinical note representations with irregular note-taking time as MISTS, such that each dimension of a series of clinical note representations is an irregular time series, and perform a time attention mechanism (Shukla & Marlin, 2021) to further model the irregularity.

Multimodal fusion. Combining both time series and clinical notes outperforms the results obtained when only one of them is used (Liu et al., 2021). Khadanga et al. (2019); Deznabi et al. (2021); Yang et al. (2021) directly concatenate representations from different modalities for downstream predictions. Yang & Wu (2021) utilizes an attention gate to fuse multimodal information. (Xu et al., 2021) selects multimodal fusion strategies from addition, concatenation and multiplication by a neural architecture search method. However, these fusion methods are only performed on EHRs without considering irregularity, failing to fully incorporate time information into multimodal representations, which is critical in real-world scenarios. To fill this gap, we first tackle irregularity in time series and clinical notes, respectively, and further leverage fusion module, which interleaves self-attentions and cross-attentions (Vaswani et al., 2017) to obtain multimodal interaction integrated with irregularity across temporal steps.

3. Method

Our method models irregularity in three portions: MISTS, clinical notes, and multimodal fusion, as shown in Figure 2. In this section, we will illustrate each part thoroughly.

3.1. Problem setup

Denote $\mathcal{D} = \{(\mathbf{x}_i^{ts}, \mathbf{t}_i^{ts}), (\mathbf{x}_i^{txt}, \mathbf{t}_i^{txt}), \mathbf{y}_i\}_{i=1}^{N}$ to be an EHR dataset with N patients, where $(\mathbf{x}_i^{ts}, \mathbf{t}_i^{ts})$ is d_m-dimensional MISTS, $\mathbf{x}_i^{ts}$ being observations and $\mathbf{t}_i^{ts}$ being corresponding time points, $(\mathbf{x}_i^{txt}, \mathbf{t}_i^{txt})$ is a series of clinical notes with note-taking time and $\mathbf{y}_i$ is the target outcome, e.g. discharge or death for modality prediction. In the following part, we drop the patient index i for simplicity. Each dimension of the MISTS, $(\mathbf{x}_j^{ts}, \mathbf{t}_j^{ts})$, where $j = 1, \cdots, d_m$, has l_j^{ts} observations, and each patient's $(\mathbf{x}^{txt}, \mathbf{t}^{txt})$ includes l^{txt} clinical notes. In early-stage medical predictions, given $(\mathbf{x}^{ts}, \mathbf{t}^{ts})$ and $(\mathbf{x}^{txt}, \mathbf{t}^{txt})$ before a certain time point (e.g. 48-hour) after admission, α, we seek to predict $\mathbf{y}$ for every patient.

3.2. MISTS

3.2.1. TDE METHODS

We will describe two TDE methods to facilitate the introduction of our proposed MISTS embedding approach. An illustration is shown in Figure 3 for better understanding.

Imputation. We first discretize $\mathbf{x}^{ts}$ based on $\mathbf{t}^{ts}$, to hourly time intervals with a sequence of regular time points, $\boldsymbol{\alpha} = [0, 1, \cdots, \alpha - 1]$. Then, for each feature, we use the last observation, if multiple observations are in the same

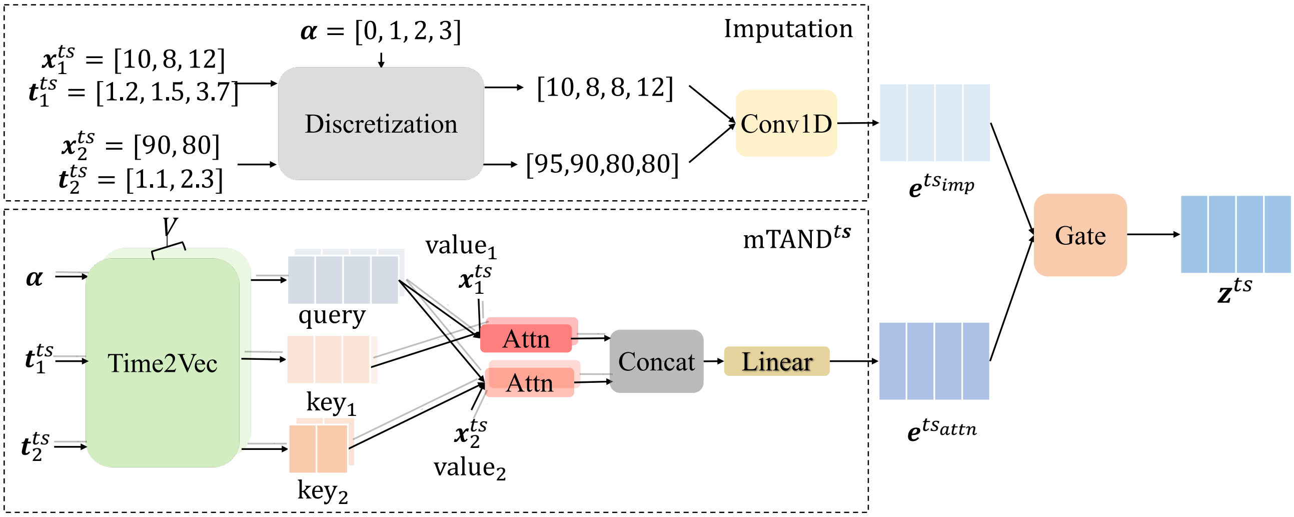

Figure 3: Architecture of UTDE module with two input features. UTDE incorporates two TDE methods: $\mathrm{Imputation}$ and mTAND^{ts}, as submodules, and learns to integrate different embeddings that are best suited to patients for a given task, via a gating mechanism.

interval, and regard intervals without any observations as missingness. We impute missing values with the most recent observation if it exists, and to the global mean of all patients otherwise. For example, with $\boldsymbol{\alpha} = [0, 1, 2, 3]$ being the first 4-hour prediction, a feature with observations $[10, 8, 12]$ collected at $[1.2, 1.5, 3.7]$ hours after admission is discretized to $[\mathrm{miss}_1, 8, \mathrm{miss}_2, 12]$, where miss_1 and miss_2 will be imputed by global mean and the previous observed value, respectively. The regular time series is fed into a 1D causal convolutional layer with stride 1 to obtain imputation embeddings with hidden dimension d_h, $\mathbf{e}^{ts_{imp}} \in \mathbb{R}^{\alpha \times d_h}$.

Discretized multi-time attention (mTAND). We leverage a discretized multi-time attention (mTAND) module (Shukla & Marlin, 2021) to re-represent MISTS into $\boldsymbol{\alpha}$.

To incorporate irregular time knowledge of MISTS, a time representation, Time2Vec (Kazemi et al., 2019), is learned to transform each value in a list of continuous time points, $\boldsymbol{\tau}$, with arbitrary length, $l_{\boldsymbol{\tau}}$, to a vector of size d_v and obtain a series of time embeddings $\theta(\boldsymbol{\tau}) \in R^{l_{\boldsymbol{\tau}} \times d_v}$,

$$\theta(\boldsymbol{\tau})[i] = \begin{cases} \omega_i \boldsymbol{\tau} + \phi_i & \text{if } i = 1 \\ sin(\omega_i \boldsymbol{\tau} + \phi_i), & \text{if } 1 < i \leq d_v, \end{cases}$$

where $\theta(\boldsymbol{\tau})[i]$ is the i-th dimension of Time2Vec, and $\{\omega_i, \phi_i\}_{i=1}^{d_v}$ are learnable parameters. The sine function captures periodic patterns while the linear term captures non-periodic behaviors, conditional on the progression of time (Kazemi et al., 2019).

The mTAND module leverages V different Time2Vec, $\{\theta_v(\cdot)\}_{v=1}^{V}$, to produce interpolation embeddings at $\boldsymbol{\alpha}$, based on a time attention mechanism. Specifically, similar to the multi-head attention (Vaswani et al., 2017), $\{\theta_v(\cdot)\}_{v=1}^{V}$ are performed on $\boldsymbol{\alpha}$ and all dimensions of MISTS to embed all time points to V different d_v-dimensional hidden spaces simultaneously, capturing various characteristics of different time points with regard to the overall time information in different time subspaces. For each $\theta_v(\cdot)$, a time attention mechanism is performed for each dimension of the MISTS simultaneously, which takes $\boldsymbol{\alpha}$ as queries, $\mathbf{t}_j^{ts}$ as keys and $\mathbf{x}_j^{ts}$ as values, and acquires $\hat{\mathbf{x}}_j^{ts} \in \mathbb{R}^{\alpha}$, a series of interpolations of corresponding univariate time series at $\boldsymbol{\alpha}$. Therefore, an interpolation matrix $\mathbf{o}_v^{ts} \in \mathbb{R}^{\alpha \times d_m}$ is obtained by

$$\begin{aligned} \mathbf{o}_v^{ts} &= [\hat{\mathbf{x}}_1^{ts}, \hat{\mathbf{x}}_2^{ts}, \cdots, \hat{\mathbf{x}}_{d_m}^{ts}] \\ \hat{\mathbf{x}}_j^{ts} &= \mathrm{Attn}(\theta_v(\boldsymbol{\alpha})\mathbf{w}_v^q, \theta_v(\mathbf{t}_j^{ts})\mathbf{w}_v^k, \mathbf{x}_j^{ts}) \end{aligned}$$

where $j = 1, \cdots, d_m$, and $\mathbf{w}_v^q$ and $\mathbf{w}_v^k$ are learned parameters. Afterwards, $\mathbf{o}_1^{ts}, \mathbf{o}_2^{ts}, \cdots, \mathbf{o}_V^{ts}$ are further concatenated and linearly projected to obtain mTAND embeddings, $\mathbf{e}^{ts_{attn}} \in \mathbb{R}^{\alpha \times d_h}$.

3.2.2. Unifying TDE methods

The imputation approach ignores the irregularity of the time series, while mTAND could result in worse performance, probably due to different time series sampling strategies (Horn et al., 2020). We propose a Unified TDE module, UTDE, via a gate mechanism to take advantage of both, for tackling complex time patterns in EHRs. The architecture of UTDE is illustrated in Figure 3. UTDE incorporates Imputation and mTAND as submodules, and learns to dynamically integrate $\mathbf{e}^{ts_{imp}}$ into $\mathbf{e}^{ts_{attn}}$ to obtain compounding embeddings $\mathbf{z}^{ts} \in \mathbb{R}^{\alpha \times d_h}$. Formally,

$$\mathbf{z}^{ts} = \mathbf{g} \odot \mathbf{e}^{ts_{imp}} + (1 - \mathbf{g}) \odot \mathbf{e}^{ts_{attn}}$$
$$\mathbf{g} = f(\mathbf{e}^{ts_{imp}} \oplus \mathbf{e}^{ts_{attn}}),$$

where $f(\cdot)$ is a gating function implemented by MLP for simplicity, $\oplus$ is the concatenation operator and $\odot$ is point-wise multiplication. Specifically, we perform UTDE in 3 levels in which $\mathbf{g}$ has different dimensions : 1) patient level with $\mathbf{g} \in \mathbb{R}$, 2) temporal level with $\mathbf{g} \in \mathbb{R}^{\alpha}$, and 3) hidden space level with $\mathbf{g} \in \mathbb{R}^{\alpha \times d_h}$. The $\mathbf{g}$ on the hidden space level can be more powerful than temporal and patient levels, while it introduces more parameters to update, making the whole module more challenging to optimize. In the experiment section, we use validation sets to decide the level on which to operate.[2] In principle, UTDE can be applied to any two TDE methods. Here, we utilize Imputation and mTAND as submodules based on empirically results.

3.3. Irregular clinical notes

To extract relevant knowledge from the clinical notes, we first encode the notes by a in-domain pretrained language model, $\mathrm{TextEncoder}$. Then we extract the representation of the [CLS] token for each encoded clinical note, to obtain a series of note representations, $\mathbf{e}^{txt} \in \mathbb{R}^{l^{txt} \times d_t}$,where d_t is the hidden dimension of the encoded text. Formally,

$$\mathbf{e}^{txt} = \mathrm{TextEncoder}(\mathbf{x}^{txt}).$$

To tackle irregularity, we sort $\mathbf{e}^{txt}$ by $\mathbf{t}^{txt}$ and cast $(\mathbf{e}^{txt}, \mathbf{t}^{txt})$ as MISTS, such that each hidden dimension of $\mathbf{e}^{txt}$ is a time series sequence and every time series sequence has the same collected time points. The mTAND module introduced in section 3.2.1 is further leveraged to re-represent $\mathbf{e}^{txt}$ into $\boldsymbol{\alpha}$. Specifically, the mTAND^{txt} takes $\boldsymbol{\alpha}$ as queries, $\mathbf{t}^{txt}$ as keys and $\mathbf{e}^{txt}$ as values and outputs $\mathbf{z}^{txt} \in \mathbb{R}^{\alpha \times d_h}$, a set of text interpolation representations at $\boldsymbol{\alpha}$. Thus we have

$$\mathbf{z}^{txt} = \mathrm{mTAND}^{txt}(\boldsymbol{\alpha}, \mathbf{t}^{txt}, \mathbf{e}^{txt}).$$

For mTAND^{ts}, the mTAND module for time series, and mTAND^{txt}, we utilize the same $\{\theta_v(\cdot)\}_{v=1}^{V}$ to encode irregular time points of two modalities to obtain temporal knowledge, because all continuous time points are in the same feature space. However, all of the other components in mTAND^{ts} and mTAND^{txt} are learned separately because the representations of time series and clinical notes are in different hidden spaces. Moreover, since the mTAND^{txt} projects $\mathbf{z}^{txt}$ to the same dimension d_h as the $\mathbf{z}^{ts}$, the dot-products are adoptable in attention modules in the fusion.

3.4. Multimodal fusion

Previous works (Khadanga et al., 2019; Deznabi et al., 2021; Yang et al., 2021; Xu et al., 2021) perform fusion strategies on multimodal data omitting irregularity. In our work, we first obtain MISTS and irregular clinical note representations, $\mathbf{z}^{ts}$ and $\mathbf{z}^{txt}$, by UTDE and mTAND^{txt}, respectively. In addition, we leverage an interleaved attention mechanism (Vaswani et al., 2017), which fuses $\mathbf{z}^{ts}$ and $\mathbf{z}^{txt}$ across temporal steps and integrates irregularity into multimodal representations, as shown in Figure 2.

Our multimodal fusion module is composed of a stack of J identical layers. Each layer consists of two self-attention sublayers and two cross-attention sublayers across temporal steps to explore the latent interactions between two modalities. Specifically, for each modality in the j-th layer, we first perform a multi-head self-attention (MH) (Vaswani et al., 2017) across temporal steps by taking the output of the corresponding modality from the $j-1$-th layer to obtain contextual embeddings. Formally, we acquire the contextual embeddings of time series and clinical notes, $\hat{\mathbf{z}}_j^{ts}$ and $\hat{\mathbf{z}}_j^{txt}$, by

$$\hat{\mathbf{z}}_j^{ts} = \mathrm{MH}_j^{ts}(\mathbf{z}_{j-1}^{ts}), \quad \hat{\mathbf{z}}_j^{txt} = \mathrm{MH}_j^{txt}(\mathbf{z}_{j-1}^{txt}),$$

where $j = 1 \dots J$, and $\mathbf{z}_0^{ts} = \mathbf{z}^{ts}$ and $\mathbf{z}_0^{txt} = \mathbf{z}^{txt}$. To capture the cross-modal information between two modalities, two multi-head cross-attentions (CMH) (Vaswani et al., 2017; Tsai et al., 2019) are leveraged to learn knowledge of another modality attended by the current modality and vice versa. Specifically, for a time series branch in the j-th layer, a CMH_j^{ts} transforms $\hat{\mathbf{z}}_j^{txt}$ to keys and values to interact with time series modality, and output $\mathbf{z}_j^{ts}$, the time series representations carrying information passed from clinical notes. For the text branch, the same process is performed but transforming $\hat{\mathbf{z}}_j^{ts}$ to keys and values, to output $\mathbf{z}_j^{txt}$, the clinical note representations integrated with information passed from time series. Formally,

$$\mathbf{z}_j^{ts} = \mathrm{CMH}_j^{ts}(\hat{\mathbf{z}}_j^{ts}, \hat{\mathbf{z}}_j^{txt}), \;\; \mathbf{z}_j^{txt} = \mathrm{CMH}_j^{txt}(\hat{\mathbf{z}}_j^{txt}, \hat{\mathbf{z}}_j^{ts}).$$

Upon the CMH output of each modality, a position-wise feedforward sublayer is stacked. We apply pre-layer normalizations and residual connections to every MH, CMH and feedforward sublayer. For simplicity, we only draw MH and CMH in multimodal fusion in Figure 2.

In this process, each modality alternately collects temporal knowledge by a MH, and updates its sequence via external information from another modality by a CMH. After $\mathbf{z}^{ts}$ and $\mathbf{z}^{txt}$ are passed through J layers, the output of each modality fully integrates information from another modality. Eventually, the last hidden states of $\mathbf{z}_J^{ts}$ and $\mathbf{z}_J^{txt}$ are extracted and concatenated to pass through a classifier with fully-connected layers to make predictions.

[2] we defer more discussion on computation resource of UTDE to Appendix A.

4. Experiments

To demonstrate the effectiveness of our methods, we conducted comprehensive experiments and ablation studies on two medical tasks: 48-hour in-hospital mortality prediction (48-IHM) and 24-hour phenotype classification (24-PHE), which are critical in the clinical scenario (Choi et al., 2016; Gupta et al., 2018).

4.1. Experimental setup

Dataset. MIMIC III is a real-world public EHR of patients admitted to ICUs, including numerical time series and clinical notes (Johnson et al., 2016). We select the MISTS features and extract clinical notes following Harutyunyan et al. (2019) and Khadanga et al. (2019), respectively. For each task, the data split of training, validation, and testing sets follows Harutyunyan et al. (2019), and patients without any clinical notes before the prediction time are removed. We defer additional data preprocessing details to the Appendix B. After preprocessing, the number of patients in the training, validation and testing sets for the 48-IHM are 11181, 2473 and 2488; and for the 24-PHE, they are 15561, 3410 and 3379, respectively.

Evaluation metric. The 48-IHM is a binary classification problem with label imbalance with death to discharge ratio of approximately 1:7. The 24-PHE is a multi-label classification problem with 25 acute care conditions, which is more changeling due to earlier prediction time and more prediction classes. We measured the performance of our proposed methods and baselines by the F1 and AUPR on 48-IHM and F1(Macro) and AUROC on 24-PHE, following the previous work (Lin et al., 2019; Arbabi et al., 2019).

MISTS baselines. We compare UTDE with a classical and 5 SOTA baselines of MISTS: Imputation, IP-Net (Shukla & Marlin, 2019), mTAND (Shukla & Marlin, 2021), GRU-D (Che et al., 2018), SeFT (Horn et al., 2020) and RAINDROP (Zhang et al., 2021b). We utilize Transformer (Vaswani et al., 2017) as backbone for UTDE and TDE methods, because Transformer has achieved SOTA results in regular time series modeling (Li et al., 2019; Lim & Zohren, 2021). We feed time series embeddings into Transformer and extract the last hidden states of the Transformer output to pass through fully-connected layers to make predictions. Following (Zhang et al., 2021b), we added two methods initially designed for forecasting tasks, DGM^2-O (Wu et al., 2021) and MTGNN (Wu et al., 2020) in our baselines. Details on MISTS baseline descriptions are in the Appendix C.1.

Irregular clinical note baselines. Considering the in-domain knowledge and the length of clinical notes, we utilize Clinical-Longformer (Li et al., 2022) with a maximum input sequence length of 1024 as our text encoder, which covers more than 98% of notes in both tasks. Same as time series modality, we feed the text interpolation representations obtained by mTAND^{txt} into Transformer for predictions. We compare our method with two baselines: T-LSTM (Baytas et al., 2017), FT-LSTM (Zhang et al., 2020), and GRU-D (Che et al., 2018), which shows strong performance in MISTS modeling. All of these methods model irregularity by acquiring a series of clinical note representations with irregular note-taking time information. To demonstrate our method's effectiveness at tackling irregularity, we further introduce two baselines: Flat (Deznabi et al., 2021), utilizing the average of clinical note embeddings of a patient for predictions, and HierTrans (Pappagari et al., 2019), utilizing Transformer to model sequential relationships among a series of clinical notes representations without considering irregular note-taking time. We defer additional baseline descriptions to the Appendix C.2.

Multimodal fusion baselines. To examine the effectiveness of our fusion method, we consider four baselines for fusion: concatenation (Khadanga et al., 2019; Deznabi et al., 2021), Tensor Fusion (Zadeh et al., 2017; Liu et al., 2018), MAG (Yang & Wu, 2021; Rahman et al., 2020a), and MulT (Tsai et al., 2019). While the first three are asynchronous methods that do not consider temporal information, MulT and our method are synchronous relying on a cross-attention mechanism to integrate information across temporal steps. Additional multimodal fusion baseline details can be found in the Appendix C.3.

4.2. Main results

In this section, we compare results between our proposed methods and their corresponding baselines in MISTS, irregular clinical notes, and multimodal fusion scenarios, respectively. The data split of each task is fixed across all methods. We conduct 3 different runs for each setting and report the corresponding mean values along with the standard deviations in testing sets, based on the best average performance on validation sets. Details for the hyperparameter selection can be found in the Appendix D.[3]

MISTS. Table 1 compares the UTDE with other time series baselines. UTDE, which incorporates two different TDE methods, obtains the best performance across two tasks on different evaluation metrics, demonstrating the advantages of our hybrid approach for downstream predictions. Specifically, UTDE relatively outperforms the strongest baseline by 4.4% in terms of AUPR on 48-IHM. Additionally, UTDE shows a 6.5% relative improvement in F1 score on the more challenging 24-PHE task compared to the best baseline. Excluding UTDE, mTAND and Imputation are the top performers on 48-IHM and 24-PHE, respectively. However, UTDE, which dynamically incorporates Imputation and mTAND, outperforms its submodules for both tasks across

[3] All experiments are conducted on 1 RTX-3090.

Table 1: Comparison between UTDE and other MISTS methods. We report average performance on three random seeds, with standard deviation as the subscript. The **Best** and 2nd best methods under each setup are bold and underlined, respectively. The performance of 48-IHM is measured on F1 and AUPR, and 24-PHE on F1 (Macro) and AUROC, respectively.

		Imputation	IP-Net	mTAND	GRU-D	SeFT	RAINDROP	DGM^2-O	MTGNN	UTDE (Ours)
48-IHM	F1	$39.73_{1.39}$	$37.22_{2.75}$	$\underline{43.87}_{0.54}$	$42.82_{0.57}$	$16.46_{8.61}$	$39.46_{3.70}$	$39.08_{1.53}$	$38.60_{2.50}$	$\mathbf{45.26}_{0.70}$
	AUPR	$44.36_{1.36}$	$39.36_{1.10}$	$\underline{47.54}_{1.28}$	$45.90_{0.40}$	$23.89_{0.46}$	$36.23_{0.37}$	$37.79_{1.54}$	$36.49_{2.10}$	$\mathbf{49.64}_{1.00}$
24-PHE	F1	$\underline{23.36}_{0.45}$	$17.90_{0.66}$	$19.90_{0.38}$	$18.96_{0.99}$	$6.10_{0.15}$	$21.81_{1.71}$	$18.40_{0.18}$	$14.48_{1.69}$	$\mathbf{24.89}_{0.43}$
	AUROC	$\underline{74.93}_{0.22}$	$73.45_{0.10}$	$73.48_{0.11}$	$73.33_{0.10}$	$65.66_{0.11}$	$73.95_{0.89}$	$71.71_{0.16}$	$70.56_{0.68}$	$\mathbf{75.56}_{0.17}$

Table 2: Results comparison in the clinical notes modality.

	48-IHM		24-PHE	
	F1	AUPR	F1	AUROC
Flat	$39.78_{1.14}$	$51.69_{0.79}$	$18.14_{1.36}$	$74.81_{0.22}$
HierTrans	$48.76_{2.44}$	$52.98_{1.69}$	$50.25_{1.21}$	$\underline{84.90}_{0.25}$
T-LSTM	$50.32_{0.89}$	$52.57_{3.25}$	$39.13_{1.35}$	$82.03_{0.07}$
FT-LSTM	$48.51_{1.67}$	$\underline{54.39}_{1.38}$	$38.24_{0.61}$	$81.07_{0.27}$
GRU-D	$\underline{51.01}_{1.50}$	$54.34_{0.75}$	$\underline{51.09}_{1.02}$	$84.19_{0.20}$
mTAND^{txt} (Ours)	$\mathbf{52.57}_{1.30}$	$\mathbf{56.05}_{1.09}$	$\mathbf{52.95}_{0.06}$	$\mathbf{85.43}_{0.07}$

Table 3: Performance comparison of different fusion strategies. Concat and TF use the concatenation and Tensor Fusion method to fuse the two modalities, respectively.

	48-IHM		24-PHE	
	F1	AUPR	F1	AUROC
TS only	$45.26_{0.70}$	$49.64_{1.00}$	$24.89_{0.43}$	$75.56_{0.17}$
Note only	$52.57_{1.30}$	$56.05_{1.09}$	$52.95_{0.06}$	$85.43_{0.07}$
Concat	$52.77_{0.70}$	$57.13_{0.7}$	$53.30_{0.35}$	$85.94_{0.21}$
TF	$51.44_{0.66}$	$57.07_{0.82}$	$49.84_{0.83}$	$84.74_{0.16}$
MAG	$53.20_{2.13}$	$57.86_{1.07}$	$53.73_{0.37}$	$85.94_{0.07}$
MulT	$\underline{54.13}_{1.20}$	$\underline{58.94}_{1.94}$	$\underline{54.20}_{0.33}$	$\underline{85.96}_{0.07}$
Interleaved (Ours)	$\mathbf{56.45}_{1.30}$	$\mathbf{60.23}_{1.54}$	$\mathbf{54.84}_{0.31}$	$\mathbf{86.06}_{0.06}$

Table 4: Ablation study on the effects of substituting different submodules in UTDE. $\mathrm{UTDE}_{\mathrm{IP-Net}}$ consists of IP-Net and Imputation, and $\mathrm{UTDE}_{\mathrm{mTAND}}$ incorporates mTAND and Imputation.

		Imputation	IP-Net	$\mathrm{UTDE}_{\mathrm{IP-Net}}$	$\mathrm{UTDE}_{\mathrm{mTAND}}$
48-IHM	F1	$39.73_{1.39}$	$37.22_{2.75}$	$\underline{44.88}_{1.96}$	$\mathbf{45.26}_{0.70}$
	AUPR	$44.36_{1.36}$	$39.36_{1.10}$	$\underline{45.49}_{3.45}$	$\mathbf{49.64}_{1.00}$
24-PHE	F1	$23.36_{0.45}$	$17.90_{0.66}$	$\underline{24.06}_{0.51}$	$\mathbf{24.89}_{0.43}$
	AUROC	$74.93_{0.22}$	$73.45_{0.10}$	$\underline{75.17}_{0.07}$	$\mathbf{75.56}_{0.17}$

various metrics, showing its ability to integrate knowledge and benefit medical predictions.

Irregular clinical notes. We compare our method with baselines in the clinical notes modality in Table 2. All of the methods that model the sequential relationships among clinical notes yield better results than Flat by a large margin, demonstrating that exploiting sequential information of clinical notes can significantly improve the downstream predictions. T-LSTM, FT-LSTM and GRU-D outperform or have comparable result compared to HierTrans on 48-IHM, but do not perform well on the more challenging 24-PHE task, where note sequences are sparser. This highlights the difficulty in modeling irregularity in sparse clinical note sequences. The proposed method, mTAND^{txt}, significantly outperforms HierTrans by relative margins of 7.8% and 5.3% in terms of F1 on the 48-IHM and 24-PHE, respectively. This shows the importance of modeling the irregularity present in clinical notes. Additionally, the results show that mTAND^{txt} surpasses other irregularity-modeling methods, particularly achieving a 3.6% relative improvement in terms of F1 on the 24-PHE, demonstrating its strong performance in tickling irregularity in clinical notes.

Multimodal fusion. We first obtain MISTS embeddings by UTDE and irregular clinical note embeddings by mTAND^{txt}, since they have the best results in each modality, and then fuse their representations via various multimodal fusion strategies. The results are shown in Table 3. Compared to models that use only one source of available data, most fusion strategies achieve better results, illustrating the effectiveness of multimodal fusion. Our fusion method yields better results than baselines for both tasks, achieving a particularly 4.3% relative improvement in F1 on the 48-IHM, showing the power of the interleaved attention mechanism. Synchronous strategies consistently achieve better results than asynchronous methods by incorporating temporal information in multimodal fusion, resulting in better integration of irregularity and fusion of different modalities. Our method further outperforms the MulT, which separately applies a cross-modal Transformer and a self-attention Transformer for each modality. This result shows that alternately obtaining temporal information and cross-modal knowledge for different modalities is more capable of fusing different modalities and integrating irregularity into multimodal representations than learning these two components separately.

4.3. Ablation study

UTDE with different submodules in MISTS. UTDE could have incorporated different TDE methods as submodules to obtain fused time series embeddings. We explored the effectiveness of the gate mechanism in UTDE by substituting mTAND to IP-Net in Table 4.

Table 5: Comparison of UTDE and its submodules with different time series backbones.

		CNN			LSTM			Transformer		
		Imputation	mTAND	UTED	Imputation	mTAND	UTED	Imputation	mTAND	UTED
48-IHM	F1	$39.66_{1.72}$	$\underline{41.40}_{1.16}$	$\mathbf{44.45}_{1.41}$	$39.72_{0.70}$	$\underline{43.61}_{0.55}$	$\mathbf{44.58}_{0.18}$	$39.73_{1.39}$	$\underline{43.87}_{0.54}$	$\mathbf{45.26}_{0.70}$
	APUR	$41.84_{0.52}$	$\underline{46.62}_{0.27}$	$\mathbf{48.22}_{0.99}$	$42.52_{0.98}$	$\underline{47.36}_{0.67}$	$\mathbf{48.17}_{0.36}$	$44.36_{1.36}$	$\underline{47.54}_{1.28}$	$\mathbf{49.64}_{1.00}$
24-PHE	F1	$\underline{20.09}_{0.70}$	$19.05_{1.17}$	$\mathbf{20.64}_{0.54}$	$19.21_{1.37}$	$\underline{19.49}_{0.32}$	$\mathbf{21.55}_{0.21}$	$\underline{23.36}_{0.45}$	$19.90_{0.38}$	$\mathbf{24.89}_{0.43}$
	AUROC	$\underline{74.69}_{0.07}$	$72.31_{0.21}$	$\mathbf{74.90}_{0.06}$	$\underline{73.95}_{0.14}$	$71.50_{0.04}$	$\mathbf{75.15}_{0.11}$	$\underline{74.93}_{0.22}$	$73.48_{0.11}$	$\mathbf{75.56}_{0.17}$

The $\mathrm{UTDE}_{\mathrm{IP-Net}}$ underperforms $\mathrm{UTDE}_{\mathrm{mTAND}}$ but still achieves better results than its submodules, Imputation and IP-Net, on both tasks, demonstrating that UTDE successfully learns from different submodules and achieves optimal performance via the gate mechanism.

UTDE with various backbones in MISTS. To evaluate the effectiveness of UTDE across different backbone encoders, we further leverage CNN (LeCun et al., 1998) and LSTM (Hochreiter & Schmidhuber, 1997) to encode time series representations obtained from TDE and UTDE methods. The results are shown in Table 5. The empirical analysis shows that Imputation and mTAND performance varies across different time series encoders. However, UTDE consistently outperforms them, demonstrating the gains of dynamically integrating different time series embeddings for medical predictions regarding the effectiveness and generalizability across time series backbones.

Does UTDE benefit performance in multimodal fusion? We drop UTDE (w/o UTDE) in our fusion model and perform only Imputation (w Imputation) and mTAND (w mTAND^{ts}) to obtain MISTS embeddings, respectively. Table 6 shows results. Consistent with the time series modality, the fusion model with learned mTAND embeddings does not consistently outperform the one with classical imputation embeddings, and vice versa. However, our fusion model with UTDE consistently surpasses those using only one TDE approach. This result further indicates that UTDE can maintain optimal performance for predictions by integrating MISTS embeddings from different TDE approaches.

Does tackling irregularity in clinical notes improve performance in multimodal fusion? We remove mTAND^{txt} and directly fuse a series of clinical notes representations with UTDE representations. The results are shown in the last row in Table 6. Performance drops when the fusion model ignores irregularity in clinical notes, showing the importance of tackling irregularity in clinical notes for medical predictions.

Table 6: Ablation study of our multimodal fusion model.

	48-IHM		24-PHE	
	F1	AUPR	F1	AUROC
Ours	$\mathbf{56.45}_{1.30}$	$\mathbf{60.23}_{1.54}$	$\mathbf{54.84}_{0.31}$	$\mathbf{86.06}_{0.06}$
:w/o UTDE				
w Imputation	$54.59_{0.91}$	$56.80_{0.54}$	$54.46_{0.17}$	$85.98_{0.02}$
w mTAND^{ts}	$54.89_{1.09}$	$59.11_{1.21}$	$54.07_{0.51}$	$85.92_{0.12}$
:w/o mTAND^{txt}	$51.14_{1.79}$	$57.81_{0.76}$	$53.33_{0.62}$	$85.60_{0.06}$

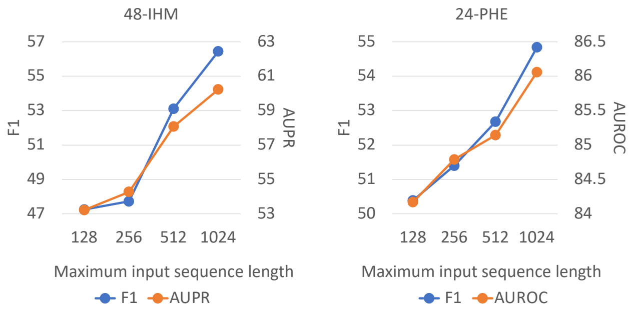

Figure 4: Performance of fusion models along with different maximum input sequence lengths.

Does the length of clinical notes affect results in multimodal fusion? Clinical notes are often lengthy and contain valuable patient information. A longer encoded clinical note brings more expressive power. We adjust our fusion model by encoding clinical notes with Bio-Clinical BERT (Alsentzer et al., 2019) with maximum input sequence lengths of 128, 256, and 512, and Clinical-Longformer (Li et al., 2022), with a maximum input sequence length of 1024, respectively. Figure 4 shows improvement in performance as maximum input sequence length increases in both tasks across various evaluation metrics, highlighting the value of clinical notes and the importance of modeling long-term dependency in text in the multimodal fusion scenario.

5. Conclusion

In this paper, we propose a unified system to fully model irregularity in multimodal EHRs for medical predictions. We first tackle irregularity in time series via a gating mechanism and long sequential clinical notes via a time attention mechanism separately, and effectively integrate irregularity into multimodal representations by an interleaved fusion strategy. We hope that our work will encourage further explorations of tackling irregularity in both single modality and multimodal scenarios.

Acknowledgments

We gratefully acknowledge financial support by the National Institutes for Health (NIH) grant NIH 7R01HL149670. We also thank all reviewers for their constructive feedback.

References

Adler-Milstein, J., DesRoches, C. M., Kralovec, P., Foster, G., Worzala, C., Charles, D., Searcy, T., and Jha, A. K. Electronic health record adoption in us hospitals: progress continues, but challenges persist. *Health affairs*, 34(12): 2174–2180, 2015.

Afessa, B., Gajic, O., and Keegan, M. T. Severity of illness and organ failure assessment in adult intensive care units. *Critical care clinics*, 23(3):639–658, 2007.

Alberti, C., Brun-Buisson, C., Burchardi, H., Martin, C., Goodman, S., Artigas, A., Sicignano, A., Palazzo, M., Moreno, R., Boulmé, R., et al. Epidemiology of sepsis and infection in icu patients from an international multicentre cohort study. *Intensive care medicine*, 28(2): 108–121, 2002.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Arbabi, A., Adams, D. R., Fidler, S., Brudno, M., et al. Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR medical informatics*, 7 (2):e12596, 2019.

Baytas, I. M., Xiao, C., Zhang, X., Wang, F., Jain, A. K., and Zhou, J. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 65–74, 2017.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):1–12, 2018.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pp. 301–318. PMLR, 2016.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Deznabi, I., Iyyer, M., and Fiterau, M. Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4026–4031, 2021.

Golmaei, S. N. and Luo, X. Deepnote-gnn: predicting hospital readmission using clinical notes and patient network. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 1–9, 2021.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

Gupta, P., Malhotra, P., Vig, L., and Shroff, G. Transfer learning for clinical time series analysis using recurrent neural networks. *arXiv preprint arXiv:1807.01705*, 2018.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9. URL https://doi.org/10.1038/s41597-019-0103-9.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Horn, M., Moor, M., Bock, C., Rieck, B., and Borgwardt, K. Set functions for time series. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4353–4363. PMLR, 13–18 Jul 2020.

Huang, K., Altosaar, J., and Ranganath, R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Kazemi, S. M., Goel, R., Eghbali, S., Ramanan, J., Sahota, J., Thakur, S., Wu, S., Smyth, C., Poupart, P., and Brubaker, M. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019.

Khadanga, S., Aggarwal, K., Joty, S., and Srivastava, J. Using clinical notes with time series data for icu management. *arXiv preprint arXiv:1909.09702*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., and Yan, X. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/6775a0635c302542da2c32aa19d86be0-Paper.pdf.

Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H., and Luo, Y. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*, 2022.

Lim, B. and Zohren, S. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.

Lin, K., Hu, Y., and Kong, G. Predicting in-hospital mortality of patients with acute kidney injury in the icu using random forest model. *International journal of medical informatics*, 125:55–61, 2019.

Lipton, Z. C., Kale, D., and Wetzel, R. Directly modeling missing data in sequences with rnns: Improved classification of clinical time series. In *Machine learning for healthcare conference*, pp. 253–270. PMLR, 2016.

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., and Morency, L.-P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018.

Liu, Z., Zhang, J., Hou, Y., Zhang, X., Li, G., and Xiang, Y. Machine learning for multimodal electronic health records-based research: Challenges and perspectives. *arXiv preprint arXiv:2111.04898*, 2021.

Mahbub, M., Srinivasan, S., Danciu, I., Peluso, A., Begoli, E., Tamang, S., and Peterson, G. D. Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult icu patients. *Plos one*, 17(1):e0262182, 2022.

McDermott, M., Nestor, B., Kim, E., Zhang, W., Goldenberg, A., Szolovits, P., and Ghassemi, M. A comprehensive ehr timeseries pre-training benchmark. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 257–278, 2021.

Otero-López, M. J., Alonso-Hernández, P., Maderuelo-Fernández, J. A., Garrido-Corro, B., Domínguez-Gil, A., and Sánchez-Rodríguez, A. Preventable adverse drug events in hospitalized patients. *Medicina clinica*, 126(3): 81–87, 2006.

Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., and Dehak, N. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 838–844. IEEE, 2019.

Rahman, W., Hasan, M. K., Lee, S., Bagher Zadeh, A., Mao, C., Morency, L.-P., and Hoque, E. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2359–2369, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.214. URL https://www.aclweb.org/anthology/2020.acl-main.214.

Rahman, W., Hasan, M. K., Lee, S., Zadeh, A., Mao, C., Morency, L.-P., and Hoque, E. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, pp. 2359. NIH Public Access, 2020b.

Rubanova, Y., Chen, R. T., and Duvenaud, D. K. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

Shukla, S. N. and Marlin, B. Interpolation-prediction networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1efr3C9Ym.

Shukla, S. N. and Marlin, B. M. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*, 2021.

Tisherman, S. A. and Stein, D. M. Icu management of trauma patients. *Critical Care Medicine*, 46(12):1991–1997, 2018.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, pp. 6558. NIH Public Access, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Wu, Y., Ni, J., Cheng, W., Zong, B., Song, D., Chen, Z., Liu, Y., Zhang, X., Chen, H., and Davidson, S. B. Dynamic gaussian mixture based deep generative model for robust forecasting on sparse multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 651–659, 2021.

Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., and Zhang, C. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 753–763, 2020.

Xiao, C., Choi, E., and Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10): 1419–1428, 2018.

Xu, Z., So, D. R., and Dai, A. M. Mufasa: Multimodal fusion architecture search for electronic health records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10532–10540, 2021.

Yang, B. and Wu, L. How to leverage multimodal ehr data for better medical predictions? *arXiv preprint arXiv:2110.15763*, 2021.

Yang, H., Kuang, L., and Xia, F. Multimodal temporal-clinical note network for mortality prediction. *Journal of Biomedical Semantics*, 12(1):1–14, 2021.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., and Eickhoff, C. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2114–2124, 2021.

Zhang, D., Thadajarassiri, J., Sen, C., and Rundensteiner, E. Time-aware transformer-based network for clinical notes series prediction. In *Machine Learning for Healthcare Conference*, pp. 566–588. PMLR, 2020.

Zhang, X., Li, S., Cheng, Z., Callcut, R., and Petzold, L. Domain adaptation for trauma mortality prediction in ehrs with feature disparity. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1145–1152, 2021a. doi: 10.1109/BIBM52615.2021.9669798.

Zhang, X., Zeman, M., Tsiligkaridis, T., and Zitnik, M. Graph-guided network for irregularly sampled multivariate time series. *arXiv preprint arXiv:2110.05357*, 2021b.

Appendix

A. Computation resource of UTDE

We set the integration level of UTDE as a hyperparameter and use validation sets to search the level on which to operate, which requires more computation resources than a model with only a single TDE method. Specifically, each time series experiment run takes less than 10 minutes with a 1 RTX-3090. The integrating operation is a hyperparameter with three levels. In this case, the total running time of UTDE will be less than 30 minutes across different integrating levels, which is affordable.

B. Data prepossessing

Table 7: Links for data generation and preprocessing used in experiments

	Links
MIMIC III	https://mimic.physionet.org/
Time series features selection and extraction	https://github.com/YerevaNN/mimic3-benchmarks
clinical notes extraction	https://github.com/kaggarwal/ClinicalNotesICU

The dataset link, and time series and clinical notes extraction used in the experiments are listed in Table 7. For time series, we follow (Harutyunyan et al., 2019) to select numerical time series features and extract time series within 48/24 hours and split the training, validation and test sets for each task. We rescale each numerical feature to be between 0 and 1. We also rescale the time to be in [0, 1] for all tasks. The clinical notes within 48/24 hours are extracted by following (Khadanga et al., 2019). For patients with more than 5 clinical notes, we utilize the last 5 clinical notes preceding the prediction time, due to computational resource limitations. We hypothesize that a note is taken closer to prediction time, the more influential it is.

Note that our early-stage phenotype classification is a brand new task compared to phenotype classification in (Harutyunyan et al., 2019), which uses the whole time series of an ICU stay. Our belief is that acute care conditions should occur during the ICU stay, and the earlier they can be predicted, the more valuable they become. Therefore, we focus on extracting the first 24 hours of data for phenotype classification, rather than using the entire admission data. This approach is also supported by Yang et al. (2021) in their research on early-stage diagnoses prediction.

C. Baselines

C.1. MISTS baselines

Imputation: Discretizes MISTS to hourly intervals and obtains imputation embeedings, as described in Section 3.2.
IP-Net (Shukla & Marlin, 2019): Employs a semi-parametric RBF interpolation network to obtain interpolation representations and a prediction network for prediction. We utilize a Transformer encoder as the prediction network.
mTAND (Shukla & Marlin, 2021): Presents a multi-time attention module to obtain an interpolation representation, as described in Section 3.2. We adopt a Transformer as the time series encoder to predict downstream tasks.
GRU-D (Che et al., 2018): Extends the GRU model to include a learnable decay term, such that the last observation is decayed to the empirical mean of time series.
SeFT (Horn et al., 2020) : Uses differentiable set function learning, such that all of the observations are first modeled individually and then pooled together via an attention based approach.
RAINDROP (Zhang et al., 2021b): Assumes that each variable of MISTS acts as a separate sensor and leverages graph neural networks to learn the dependencies between different variables.
DGM^2-O (Wu et al., 2021): A model initially designed for forecasting tasks, that utilizes a kernel-based approach to interpolate irregular time series.
MTGNN (Wu et al., 2020): A graph neural network initially designed for forecasting tasks, in which the inter-variate relationships are constructed by connecting each node with its top k nearest neighbors in a defined metric space.
The implementations of IP-Net (Shukla & Marlin, 2019) and mTAND (Shukla & Marlin, 2021) follow the original paper[4] [5]. We directly adopt the implementations of GRU-D (Che et al., 2018), SeFT (Horn et al., 2020), RAINDROP (Zhang et al.,

[4] https://github.com/mlds-lab/interp-net

[5] https://github.com/reml-lab/mTAN

2021b), DGM^2-O (Wu et al., 2021) and MTGNN (Wu et al., 2020) provided by (Zhang et al., 2021b) [6].
Following (Zhang et al., 2021b), predictions with forecasting models are designed as single-step forecasting problems.

C.2. Irregular clinical notes baselines

Time-Aware LSTM (T-LSTM) (Baytas et al., 2017): A variant of LSTM taking the elapsed time between notes into account with a decreasing function.
Flexible Time-aware LSTM (FT-LSTM) (Zhang et al., 2020): Encodes the temporal information of clinical notes by utilizing time-aware trainable parameters in an LSTM cell.
We utilize Clinical-Longformer with a maximum sequence length of 1024 (Li et al., 2022) as the text encoder by using the pre-trained weights provided in HuggingFace (Wolf et al., 2020)[7]. We directly adopt the implementations of T-LSTM and FT-LSTM provided by (Zhang et al., 2020). and GRU-D (Che et al., 2018) provided by (Zhang et al., 2021b). We leverage the same implementation of mTAND as MISTS baseline.

C.3. multimodal fusion baselines

Multimodal Adaptation Gate (MAG) (Rahman et al., 2020b; Yang & Wu, 2021):Adjusts the representation of one modality with a displacement vector derived from the other modalities.
Tensor Fusion (TF) (Zadeh et al., 2017; Liu et al., 2018): Performs an outer product on representations of different modalities.
Multimodal Transformer (MulT) (Tsai et al., 2019): Uses a cross-modal Transformer followed by a self-attention Transformer to obtain multimodal representations across time steps for each modality.
We utilize the implementations of MAG and TF provided by (Yang et al., 2021) [8], and MulT (Tsai et al., 2019) provided by the original paper[9]. We perform Concat, MAG and TF as late fusion by first applying a Transformer on every modality to acquire representations of different modalities, and then integrating the last hidden state of every single modality with different fusion strategies to obtain multimodal representations for downstream tasks.

D. Hyperparameters and training details

We use a batch size of 32 and learning rate for pre-trained language models (PLMs) of 2×10^{-5} and others of 0.0004. We use the Adam algorithm for gradient-based optimization (Kingma & Ba, 2014). We store the parameters that obtain the highest F1 and Macro-F1 in the validation set, and use it to make predictions for testing samples for 48-IHM and 24-PHE, respectively. The chosen hyperparameters are the same across tasks (48-IHM and 24-PHE) and models (both baselines and our methods) based on MISTS, irregular clinical note and multimodal fusion settings.

D.1. MISTS

For all MISTS models, we run the models for 20 epochs. We search for hidden units of Imputation, mTAND, IP-Net, GRU-D and SeFT, over the range {64,128}. For Imputation, we set the kernel size of 1D Convolution as 1. For mTAND we search for hidden size of time embeddings over the range {64,128} and take the the number of time embeddings, V, to be 8. We utilize a 3-layer Transformer as the backbone encoder for Imputation, mTAND and IP-Net. For UTDE, we search the hyperparameters of submodules Imputation and mTAND over the same range as the model with only a single method, and use a 3-layer Transformer as backbone encoder. We search for the gate integration level in {"patient", "temporal", "hidden space" }.

D.2. Irregular clinical notes

In our primary study, we empirically found that all models in the clinical note modality converge within 6 epochs, so that we train all the models for 6 epochs. In addition, we found that fine-tuning the PLM in the first 3 epochs and regarding the PLM as a feature extractor in later epochs achieved better results than fine-tuning the PLM in the whole training. We search for

[6] https://github.com/mims-harvard/Raindrop
[7] https://huggingface.co/yikuan8/Clinical-Longformer
[8] https://github.com/emnlp-mimic/mimic
[9] https://github.com/yaohungt/Multimodal-Transformer

hidden units of T-LSTM, FT-LSTM, GRU-D and mTAND^{txt} over the range $\{$64,128$\}$. For mTAND^{txt}, time embeddings hidden size is searched over the range $\{$64,128$\}$ and the number of embeddings V is equal to 8.

D.3. Multimodal fusion

. Same as the clinical note modality, we run all fusion models for 6 epochs, and fine-tune the PLM in the first 3 epochs. We utilize 3-layer Transformer encoders to encode each modality for Concat, MAG and TF. For MulT, we perform 3 layer cross-modal Transformer followed by a 3 layer self-attention Transformer for each modality. We learn a 3 layer interleaved Transformer for our multimodal fusion strategy (J=3). We search for the hyperparameters of UTDE and mTAND^{txt} over the same range in each single modality setting. We search for the hidden size of Transformers over the range $\{$64,128$\}$.