Pseudo Code for TimeCHEAT

Algorithm 1: TimeCHEAT Algorithm

Input: Train finite sequence of pairs $S = (\widetilde{X}, Y)$, where $\widetilde{X} \in \mathbb{R}^{N \times C \times T}$, the number of network layers L, Patch number PN, correlation matrix **CM**, the number of reference points T_p.

Parameters: Embedding model g, encoder model $f_{\text{enc}}^{(i)}, i = 1, \ldots, C$, decoder model f_{dec}.

Output: Embedding model g, encoder model $f_{\text{enc}}^{(i)}, i = 1, \ldots, C$, decoder model f_{dec}, correlation matrix **CM**.

Splite $\widetilde{X}$ into PN patches, i.e., $\left\{\widetilde{X}_p\right\}_{p=1}^{PN}$
for $p \leftarrow 1$ to PN **do**
  $V_C, V_T, V_\tau, E \leftarrow \text{I2RGraph}^{-1}(\widetilde{X}_p)$
  // Transfer ISMTS data into a graph structure
  $h_c^{\text{node},0} \leftarrow \textbf{FFN}\left(\textbf{CM}\left(c\right)\right), c \in V_C$
  $h_t^{\text{node},0} \leftarrow \sin\left(\textbf{FFN}\left(t\right)\right), t \in V_T \dot{\cup} V_{\boldsymbol{\tau}}$
  $h_e^{\text{edge},0} \leftarrow \textbf{FFN}\left(ei\right), e \in E$
  // Graph Neural Network
  for $l \leftarrow 1$ to L **do**
    for $u \in V$ **do**
      $h_u^{\text{node},l+1} \leftarrow \mathbf{MultiHead}^{(l)}\left(h_u^{\text{node},l}, H_u, H_u\right)$
      $s.t. \quad H_u \leftarrow \left(\left[h_v^{\text{node},l} \| h_e^{\text{edge},l}\right]\right)_{v \in \mathcal{N}(u)}$
    end for
    for $e = \{u, v\} \in E$ **do**
      $h_e^{\text{edge},l+1} \leftarrow \alpha\left(h_e^{\text{edge},l} + \mathbf{FFN}^{(l)}\left(h_c^{\text{node},l} \left\| h_t^{\text{node},l} \right\| h_e^{\text{edge},l}\right)\right)$
    end for
  end for
  $H_p \leftarrow \text{I2RGraph}(h^{\text{node},L}, h^{\text{edge},L})$
end for
$H = [H_1 \| \ldots \| H_P] \in \mathbb{R}^{N \times PN \times T_p \times C}$
Permute dimensions: $\widetilde{H} \in \mathbb{R}^{N \times C \times PN \times T_p}$
// Representation Learning
$H^{PE} \leftarrow \widetilde{H} + PE$ // PE denotes Position Embedding
for $c \leftarrow 1$ to C **do**
  $R_c \leftarrow f_{\text{enc}}^{(c)}\left(H_c^{PE}\right)$
end for
$Y \leftarrow f_{dec}\left(R\right)$

Further Details on Datasets

We adopt the data processing approach used in RAINDROP (Zhang et al. 2021) for the classification task, mTANs (Shukla and Marlin 2021) for the interpolation task, and GraFITi (Yalavarthi et al. 2024) for the forecasting task. The aforementioned processing methods serve as the usual setup, which our method also follows for fair comparison. *However, it's important to note that we do not incorporate static attribute vectors* (such as age, gender, time from hospital to ICU admission, ICU type, and length of stay in ICU) in our processing. This decision is based on the fact that our model, TimeCHEAT, is not specifically designed for clinical datasets. Instead, it is designed as a versatile, general model capable of handling various types of datasets, which may not always include such static vectors. The detailed information of baselines is in Table 5. *We can see that the high missing ratios in most real-world ISMTS make their analysis particularly challenging.*

Datasets for Classification

P19: PhysioNet Sepsis Early Prediction Challenge 2019. P19 dataset (Reyna et al. 2020) comprises data from $38,803$ patients, each monitored by 34 irregularly sampled sensors, including 8 vital signs and 26 laboratory values. The original dataset contained $40,336$ patients, but we excluded those with excessively short or long time series, resulting in a range of 1 to 60 observations per patient as in RAINDROP. Each patient has a binary label representing the occurrence of sepsis within the next 6 hours. The dataset has a high imbalance with approximately $\sim 4\%$ positive samples.

P12: PhysioNet Mortality Prediction Challenge 2012. P12 (Goldberger et al. 2000) includes data from $11,988$ patients after removing inappropriate 12 samples as explained in (Horn et al. 2020). This dataset features multivariate time series from 36 sensors collected during the first 48 hours of ICU stay. Each patient has a binary label indicating the length of stay in the ICU, in which a negative label for stays under 3 days and a positive label for longer stays. P12 is imbalanced with $\sim 93\%$ positive samples.

PAM: PAMAP2 Physical Activity Monitoring. PAM (Reiss and Stricker 2012) records the daily activities of 9 subjects using 3 inertial measurement units. RAINDROP has adapted it for irregularly sampled time series classification by excluding the ninth subject for short sensor data length. The continuous signals were segmented into samples with the window size 600 and 50% overlapping rate. Originally with 18 activities, we retain 8 with over 500 samples each, while others are dropped. After modification, PAM includes $5,333$ sensory signal segments, each with 600 observations from 17 sensors at 100 Hz. To simulate irregularity, 60% of observations are randomly removed by RAINDROP, uniformly across all experimental setups for fair comparison. The 8 classes of PAM represent different daily activities, with no static attributes and roughly balanced distribution.

Dataset for Interpolation

Physionet: PhysioNet Challenge 2012 dataset. Physionet (Reiss and Stricker 2012) comprises 37 variables from ICU patient records, with each record containing data from the first 48 hours after admission to ICU. Aligning with the methodology of Neural ODE (Rubanova, Chen, and Duvenaud 2019), we round observation times to the nearest minute, resulting in up to $2,880$ potential measurement times for each time series. The dataset encompasses $4,000$ labeled instances and an equal number of unlabeled instances. For our study, we utilize all $8,000$ instances in interpolation experiments. Our primary objective is to predict in-hospital mortality, with 13.8% of the instances belonging to the positive class.

Table 5: Statistics of the ISMTS datasets used in our experiments. "#Avg. obs." denotes the average number of observations for each sample.

Tasks	Datasets	#Samples	#Variables	#Avg. obs.	#Classes	Imbalanced	Missing ratio
Classification	P19	38,803	34	401	2	True	94.9%
	P12	11,988	36	233	2	True	88.4%
	PAM	5,333	17	4,048	8	False	60.0%
Interpolation	PhysioNet	4,000	37	2,880	-	-	78.0%
Forecasting	USHCN	1,100	5	263	-	-	77.9%
	MIMIC-III	21,000	96	274	-	-	94.2%
	MIMIC-IV	18,000	102	496	-	-	**97.8**%
	Physionet12	5,333	37	130	-	-	85.7%

Dataset for Forecasting

USHCN: U.S. Historical Climatology Network. USHCN (Menne, Williams Jr, and Vose 2015) data are used to quantify national and regional-scale temperature changes in the contiguous United States. It contains measurements of 5 variables from 1280 weather stations. Following the preprocessing proposed by (De Brouwer et al. 2019), the majority of the over 150 years of observations are excluded, and only data from the years 1996 to 2000 are used in the experiments. Furthermore, to create a sparse dataset, only a randomly sampled 5% of the measurements are retained.

Physionet12. This dataset consists of medical records from $12,000$ ICU patients. During the first 48 hours of admission, measurements of 37 vital signs were recorded. Following the forecasting approach used in recent work, such as (Yalavarthi et al. 2024; Biloš et al. 2021; De Brouwer et al. 2019), we pre-process the dataset to create hourly observations, resulting in a maximum of 48 observations per series.

MIMIC-III & MIMIC-IV: Medical Information Mart for Intensive Care. MIMIC-III (Johnson et al. 2016) and MIMIC-IV (Johnson et al. 2020) are widely used medical datasets providing critical insights into ICU patient care. MIMIC-IV builds on MIMIC-III's success, featuring updated data and a modular organization that emphasizes data provenance and supports diverse healthcare applications. To capture a broad spectrum of patient characteristics and conditions, MIMIC-III tracks 96 variables, while MIMIC-IV tracks 102. We followed standard preprocessing steps from previous studies(Yalavarthi et al. 2024; Schirmer et al. 2022; Biloš et al. 2021; De Brouwer et al. 2019), rounding observations to 30-minute intervals and using only the first 48 hours post-admission. Patients with ICU stays shorter than 48 hours were excluded from the analysis.

Further Ablation Experimental Results

We take the classification task as an example to conduct the ablation study. We verify the necessity of three main designs in TimeCHEAT: 1) learnable correlation between multiple channels in the embedding procedure, 2) channel-dependent embedding learning without special assumptions, and 3) channel-independent Transformer encoder.

Therefore, our ablation study contains the following 4 conditions.

1. **w/o correlation:** remove the learnable correlation and replace it with onehot binary indicator vector,
2. **w/ iTransformer:** use channel-dependent Transformer encoder,
3. **w/o correlation + w/ iTransformer:** use onehot binary indicator vector and channel-dependent Transformer encoder,
4. **mTAND instead:** replace the channel-dependent bipartite graph with channel-independent mTAND for embedding learning, therefore it can also be seen as a channel-independent method.

As shown in Table 6, the full TimeCHEAT framework, which includes all original components (line 7), delivers the best performance. When local correlations between channels are removed (line 3), by replacing the channel encoding with a one-hot binary indicator vector (resulting in CI embedding learning), sparse sampling channels lack sufficient information and fail to aggregate crucial data from related channels for improved embedding. Discarding the CI encoder and switching to a vanilla Transformer (line 4) leads to a significant drop in classification accuracy, underscoring the effectiveness of the CI strategy in the encoding phase. Combining the above two changes results in a fully CI model (line 5), which yields nearly the worst accuracy among all tested conditions.

Finally, we conducted an additional experiment by replacing the local graph embedding with CI mTAND (line 5) to evaluate the effectiveness of the I2RGraph. While using mTAND within a patch can partially address issues related to its assumptions about timestamp distances, it still falls short in capturing the correlation between channels. We applied mTAND locally for embedding learning and a Transformer globally to model long-range relationships, which helps avoid overlooking long-range correlations when mTAND is used globally with timestamp distances as weights. However, this model underperforms because mTAND does not effectively capture inter-channel correlations, effectively making it a CI strategy.

Methods	P19		P12		PAM			
	AUROC	AUPRC	AUROC	AUPRC	Accuracy	Precision	Recall	F1 score
w/o correlation	88.0 ± 3.1	54.4 ± 5.0	83.5 ± 0.9	46.5 ± 2.3	94.6 ± 1.0	95.8 ± 0.7	95.3 ± 0.9	95.5 ± 0.8
w/ iTransformer	87.6 ± 2.4	54.7 ± 4.6	79.9 ± 1.5	39.2 ± 3.2	93.1 ± 1.2	94.2 ± 1.1	94.0 ± 1.0	94.1 ± 0.9
w/o correlation + w/ iTransformer	86.8 ± 2.7	54.2 ± 4.3	80.1 ± 1.7	39.4 ± 3.1	93.0 ± 0.9	94.1 ± 0.8	94.0 ± 0.6	94.0 ± 0.7
mTAND instead	87.4 ± 2.3	52.5 ± 3.4	84.3 ± 0.8	48.2 ± 1.0	95.8 ± 0.8	96.4 ± 0.9	96.1 ± 0.5	96.6 ± 0.6
TimeCHEAT	$\mathbf{89.5} \pm 1.9$	$\mathbf{56.1} \pm 4.6$	$\mathbf{84.5} \pm 0.7$	$\mathbf{48.2} \pm 1.9$	$\mathbf{96.5} \pm 0.6$	$\mathbf{97.1} \pm 0.5$	$\mathbf{96.9} \pm 0.6$	$\mathbf{97.0} \pm 0.5$

Table 6: Full ablation studies on different strategies of TimeCHEAT in classification.

Experimental details

TimeCHEAT parameters

We present the training hyperparameters and model parameters here. The maximum epoch is set to 200, and AdamW optimizer is selected as our optimizer without weight decay. By default, the learning rate is set to $1e-3$, and the learning rate schedule is OnPlateau, which reduces learning rate when a metric has stopped improving. Batch size for all datasets is set to 50, and the number of layers and the latent dimension in I2RGraph are set to 2 and 128, respectively. We set the reference point number to 16 for each patch on all datasets, and the number of patches is set to 8. Transformer contains 3 encoder layers with head number $H = 8$, and dimension of latent space $D = 256$. The feed forward network in the Transformer encoder block consists of 2 linear layers with GELU (Nie et al. 2023) activation function.

All the models were experimented with using the PyTorch library on a GeForce RTX-3090 GPU.

Baseline Parameters

The implementation of baseline models adheres closely to the methodologies outlined in their respective papers, including SeFT (Horn et al. 2020), GRU-D (Che et al. 2018), mTAND (Shukla and Marlin 2021) and ViTST (Li, Li, and Yan 2023).

References

Biloš, M.; Sommer, J.; Rangapuram, S. S.; Januschowski, T.; and Günnemann, S. 2021. Neural flows: Efficient alternative to neural ODEs. *NeurIPS*, 34: 21325–21337.

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

De Brouwer, E.; Simm, J.; Arany, A.; and Moreau, Y. 2019. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. *NeurIPS*, 32.

Goldberger, A. L.; Amaral, L. A.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C.-K.; and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220.

Horn, M.; Moor, M.; Bock, C.; Rieck, B.; and Borgwardt, K. 2020. Set functions for time series. In *ICML*, 4353–4363. PMLR.

Johnson, A.; Bulgarelli, L.; Pollard, T.; Horng, S.; Celi, L. A.; and Mark, R. 2020. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, 49–55.

Johnson, A.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database Sci. *Data*, 3(1): 1.

Li, Z.; Li, S.; and Yan, X. 2023. Time Series as Images: Vision Transformer for Irregularly Sampled Time Series. In *NeurIPS*.

Menne, M. J.; Williams Jr, C.; and Vose, R. S. 2015. United States historical climatology network daily temperature, precipitation, and snow data. *Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee*.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *ICLR*.

Reiss, A.; and Stricker, D. 2012. Introducing a new benchmarked dataset for activity monitoring. In *16th international symposium on wearable computers*, 108–109. IEEE.

Reyna, M. A.; Josef, C. S.; Jeter, R.; Shashikumar, S. P.; Westover, M. B.; Nemati, S.; Clifford, G. D.; and Sharma, A. 2020. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Critical care medicine*, 48(2): 210–217.

Rubanova, Y.; Chen, R. T.; and Duvenaud, D. K. 2019. Latent ordinary differential equations for irregularly-sampled time series. *NeurIPS*, 32.

Schirmer, M.; Eltayeb, M.; Lessmann, S.; and Rudolph, M. 2022. Modeling irregular time series with continuous recurrent units. In *ICML*, 19388–19405. PMLR.

Shukla, S. N.; and Marlin, B. 2021. Multi-Time Attention Networks for Irregularly Sampled Time Series. In *ICLR*.

Yalavarthi, V. K.; Madhusudhanan, K.; Scholz, R.; Ahmed, N.; Burchert, J.; Jawed, S.; Born, S.; and Schmidt-Thieme, L. 2024. GraFITi: Graphs for Forecasting Irregularly Sampled Time Series. In *AAAI*, 16255–16263.

Zhang, X.; Zeman, M.; Tsiligkaridis, T.; and Zitnik, M. 2021. Graph-Guided Network for Irregularly Sampled Multivariate Time Series. In *ICLR*.