TE-ESN: Time Encoding Echo State Network for Prediction Based on Irregularly Sampled Time Series Data

Chenxi Sun[1,2], **Shenda Hong**[3,4], **Moxian Song**[1,2], **Yen-Hsiu Chou**[1,2], **Yongyue Sun**[1,2], **Derun Cai**[1,2] and **Hongyan Li**[1,2*]
[1]Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China.
[2]School of Electronics Engineering and Computer Science, Peking University, Beijing, China.
[3]National Institute of Health Data Science, Peking University, Beijing, China.
[4]Institute of Medical Technology, Health Science Center of Peking University, Beijing, China.
{sun_chenxi,hongshenda,songmoxian,emilychou,redhated,cdr,leehy}@pku.edu.cn

Abstract

Prediction based on Irregularly Sampled Time Series (ISTS) is of wide concern in real-world applications. For more accurate prediction, methods had better grasp more data characteristics. Different from ordinary time series, ISTS is characterized by irregular time intervals of intra-series and different sampling rates of inter-series. However, existing methods have suboptimal predictions due to artificially introducing new dependencies in a time series and biasedly learning relations among time series when modeling these two characteristics. In this work, we propose a novel Time Encoding (TE) mechanism. TE can embed the time information as time vectors in the complex domain. It has the properties of absolute distance and relative distance under different sampling rates, which helps to represent two irregularities. Meanwhile, we create a new model named Time Encoding Echo State Network (TE-ESN). It is the first ESNs-based model that can process ISTS data. Besides, TE-ESN incorporates long short-term memories and series fusion to grasp horizontal and vertical relations. Experiments on one chaos system and three real-world datasets show that TE-ESN performs better than all baselines and has better reservoir property.

1 Introduction

Prediction based on Time Series (TS) widely exists in many scenarios, such as healthcare and meteorology [Xing *et al.*, 2010; Wang *et al.*, 2019]. Many methods, especially Recurrent Neural Networks (RNNs), have achieved state-of-the-art [Fawaz *et al.*, 2019]. However, in real-world applications, TS usually is Irregularly Sampled Time Series (ISTS) data. For example, the blood sample of a patient during hospitalization is not collected at a fixed time of day or week. This characteristic limits the performances of the most methods.

Comprehensive learning of ISTS characteristics contributes to the accuracy of final prediction [Hao and Cao, 2020]. ISTS has two characteristics of irregularity under the aspects of intra-series and inter-series:

- Intra-series irregularity is the irregular time intervals between observations within a time series. For example, due to the change of patient's health status, the relevant measurement requirements are also changing. In Figure 1, the time between a COVID-19 patient's blood sample could be 1 hour or even 7 days. Uneven intervals will change the dependency between observations and large time intervals will add a time sparsity factor [Jinsung *et al.*, 2017].
- Inter-series irregularity is the different sampling rates among time series. For example, in Figure 1, because vital signs have different rhythms and sensors have different sampling time, for a COVID-19 patient, heart rate is measured in seconds, while blood sample is collected in days. The difference of sampling rates is not conducive to data preprocessing and model design [Karim *et al.*, 2019].

However, grasping both two irregularities is challenging. In real-world applications, a model usually has multiple time series as input. If seeing the input as a multivariate time series, data alignment with up/down-sampling and imputation occur. But it will artificially introduce some new dependencies while omit some original dependencies, causing suboptimal prediction [Sun *et al.*, 2020b]; If seeing the input as multiple separated time series and changing dependencies based on time intervals, the method will encounter the problem of bias, embedding stronger short-term dependency in high sampled time series due to smaller time intervals. This is not necessarily the case, for example, although the detection of blood pressure is not frequent than heart rate in clinical practice, its values have a strong diurnal correlation [Virk, 2006].

In order to get rid of the above dilemmas and achieve more accurate prediction, modeling all irregularities without introducing new dependency is feasible. However, the premise is that ISTS can't be interpolated, which makes the alignment impossible, leading to batch gradient descent for multivariate time series hard to implement, aggravating the non-converging and instability of error Back Propagation RNNs (BPRNNs), the basis of existing methods for ISTS [Sun *et al.*, 2020b]. Echo State Networks (ESNs) is a simple type of RNNs and can avoid non-converging and computationally

*Contact Author. Peking University, No. 5 Yiheyuan Road, Beijing 100871, People's Republic of China.

expensive by applying least square problem as the alternative training method [Jaeger, 2002]. But ESNs can only process uniform TS by assuming time intervals are equally distributed, with no mechanism to model ISTS. For solving all the difficulties mentioned above, we design a new structure to enable ESNs to handle ISTS data, where a novel mechanism makes up for the disadvantage of no learning of irregularity.

- We introduce a novel mechanism named Time Encoding (TE). TE represents time points as dense vectors and extends to complex domain for more options. TE injects the absolute and relative distance properties based on time interval and sampling rate into time representations, which helps model both intra-series irregularity and inter-series irregularity of ISTS at the same time.
- We design a mode named Time Encoding Encoding Echo State Network (TE-ESN). In addition to the ability of modeling both two ISTS irregularities, TE-ESN can learn the long short-term memories in a time series longitudinally and fuses the relations among time series horizontally.
- We evaluate TE-ESN for early prediction and one-step-ahead forecasting on four datasets. TE-ESN outperforms state-of-the-art models and has better reservoir property.

2 Related Work

2.1 ISTS Method Categories

Existing methods can be divided into two categories:

Missing data perspective. It discretizes the time axis into non-overlapping intervals, points without data are considered as missing data. M-RNN [Jinsung *et al.*, 2017] handled missing data by operating time series forward and backward. GRU-D [Che *et al.*, 2018] used decay rate to weigh the correlation between missing data and other data. But data imputation may artificially introduce new dependency beyond original relations and totally ignore ISTS irregularities.

Raw data perspective. It constructs models which can directly receive ISTS as input. T-LSTM [Baytas *et al.*, 2017] used the elapsed time function for modeling irregular time intervals. IPN [Shukla and Marlin, 2019] used three time perspectives for modeling different sampling rates. However, they just performed well in the univariate time series, for multiple time series, they had to apply alignment first, causing the data missing in some time points, back to the defects of the first category.

2.2 Echo State Networks

The adaption of the BPRNNs training requirements causes the above defects. ESNs with a strong theoretical ground, is practical and easy to implement, can avoid non-converging [Gallicchio and Micheli, 2017; Sun *et al.*, 2020a]. [Jaeger *et al.*, 2007] designed a classical reservoir structure leaky-ESN using leaky integrator neurons and mitigated noise problem. [Gallicchio *et al.*, 2017] proposed a stacked reservoirs structure DeepESN based on deep learning (DL) to pursue conciseness of ESNs and effectiveness of DL. [Zheng *et al.*, 2020] proposed LS-ESN by considering the relations of time series in different time spans. But there is no ESNs-based methods for ISTS.

3 Time Encoding Echo State Network

The widely used RNN-based methods, especially ESNs, only model the order of time series by assuming the time distribution is uniform. We design Time Encoding (TE) mechanism (Section 3.2) to help ESNs model ISTS (Section 3.3).

3.1 Definitions

First, we give two new definitions used in this paper.

Definition 1 (Irregularly Sampled Time Series ISTS)**.** *A time series u with sampling rate $r_s(d), d \in= \{1, ..., D\}$ has several observations distributed with time $t, t \in= \{1, ..., T\}$. u_t^d represents an observation of a time series with sampling rate $r_s(d)$ in time t.*

ISTS has two irregularities: (1) Irregular time intervals of intra-series: $t_i - t_{i-1} \neq t_j - t_{j-1}$. (2) Different sampling rate of inter-series: $r_s(d_i) \neq r_s(d_j)$.

For prediction tasks, one-step-ahead forecasting is using the observed data $u_{1:t}$ to predict the value of u_{t+1}, and continues over time; Early prediction is using the observed data $u_{1:t}$ $(t < t_{pre})$ to predict the classes or values in time t_{pre}.

Definition 2 (Time Encoding TE)**.** *Time encoding mechanism aims to design methods to embed and represent every time point information of a time line.*

TE mechanism extends the idea of Positional Encoding (PE) in natural language processing. PE was first introduced to represent word positions in a sentence [Gehring *et al.*, 2017]. Transformer [Vaswani *et al.*, 2017] model used a set of sinusoidal functions discretized by each relative input position, shown in Equation 1. Where pos indicates the position of a word, d_{model} is the embedding dimension. Meanwhile, a recent study [Wang *et al.*, 2020] encoded word order in complex embeddings. An indexed j word in the pos position is embeded as $g_{pe}(j, pos) = re^{i\omega_j pos + \theta_j}$. r, ω and θ denote amplitude, frequency and primary phase respectively. They are all the parameters that should to be learned using deep learning model.

$$\begin{cases} PE(pos, 2i) = \sin(\dfrac{pos}{10000^{\frac{2i}{d_{model}}}}) \\ PE(pos, 2i+1) = \cos(\dfrac{pos}{10000^{\frac{2i}{d_{model}}}}) \end{cases} \tag{1}$$

3.2 Time Encoding Mechanism

First, we introduce how the Time Vector (TV) perceives irregular time intervals of a single ISTS with the fixed sampling rate. Then, we show how Time Encoding (TE) embeds time information of multiple ISTS with different sampling rates. The properties and proofs are summarized in the Appendix.

Time Vector with Fixed Sampling Rate

Now, let's only consider one time series, whose irregularity is just reflected in the irregular time intervals. Inspired by Positional Encoding (PE) in Equation 1, we apply Time Vector (TV) to note the time codes. Thus, in a time series, each time point is tagged with a time vector:

$$\begin{aligned} TV(t) &= [..., \sin(c_i t), \cos(c_i t), ...] \\ c_i &= MT^{-\frac{2i}{d_{TV}}}, i = 0, ..., \frac{d_{TV}}{2} - 1 \end{aligned} \tag{2}$$

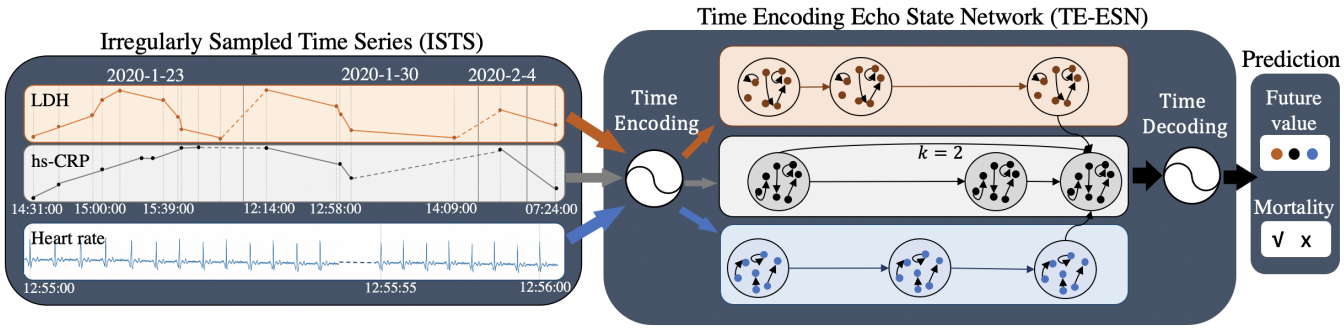

Figure 1: An ISTS example of a COVID-19 patient and the structure of TE-ESN

In Equation 2, each time vector has d_{TV} embedding dimensions. Each dimension corresponds to a sinusoidal. Each sinusoidal wave forms a geometric progression from 2π to $MT\pi$. MT is the biggest wavelength defined by the maximum number of input time points.

Without considering the different sampling rates of inter-series, for a single ISTS, TV can simulate the time intervals between two observations by its properties of absolute distance and relative distance.

Property 1 (Absolute Distance Property)**.** *For two time points with distance p, the time vector in time point $t+p$ is the linear combination of the time vector in time point t.*

$$\begin{aligned} TV(t+p) &= (a,b) \cdot TV(t) \\ a = TV(p, 2i+1), &b = TV(p, 2i) \end{aligned} \tag{3}$$

Property 2 (Relative Distance Property)**.** *The product of time vectors of two time points t and $t+p$ is negatively correlated with their distance p. The larger the interval, the smaller the product, the smaller the correlation.*

$$TV(t) \cdot TV(t+p) = \sum_{i=0}^{\frac{d_{TV}}{2}-1} \cos(c_i p) \tag{4}$$

For a computing model, if its inputs have the time vectors of time points corresponding to each observation, then the calculation of addition and multiplication within the model will take the characteristics of different time intervals into account through the above two properties, improving the recognition of long term and short term dependencies of ISTS. Meanwhile, without imputing new data, natural relation and dependency within ISTS are more likely to be learned.

Time Encoding with Different Sampling Rates

When the input is multi-series, another irregularity of ISTS, different sampling rates, shows up. Using the above introduced time vector will encounter the problem of bias. It will embed more associations between observations with high sampling rate according to the Property 2, as they have smaller time intervals. But we can not simply conclude that the correlation between the values of time series with low sampling rate is weak.

Thus, we design an advanced version of time vector, noted Time Encoding (TE), to encode time within multiple ISTS. TE extends TV to complex-valued domain. For a time point t in the d-th ISTS with $r_s(d)$ sampling rate, the time code is in Equation 5, where ω is the frequency.

$$TE(d,t) = e^{i(\omega t)}, \omega = \omega_d \cdot r_s^{-1}(d) \tag{5}$$

Compared with TV, TE has two advantages:

The first is that TE not only keeps the property 1 and 2, but also incorporates the influence of frequency ω, making time codes consistent at different sampling rates.

ω reflects the sensitivity of observation to time, where a large ω leads to more frequent changes of time codes and more difference between the representations of adjacent time points. For relative distance property, a large ω makes the product large when distance p is fixed.

Property 3 (Relative Distance Property with ω)**.** *The product of time encoding of two time points t and $t+p$ is positive correlated with frequency ω.*

$$TE(t) \cdot TE(t+p) = e^{i\omega t} \cdot e^{i\omega(t+p)} = e^{i\omega(2t+p)} \tag{6}$$

In TE, we set $\omega = \omega_d \cdot r_s^{-1}(d)$. ω_d is the frequency parameter of d-th sampling rate. TE fuses the sampling rate term $r_s^{-1}(d)$ to avoid the bias of time vector causing by only considering the effect of distance p.

The second is that each time point can be embeded into d_{TE} dimensions with more options of frequencies by setting different $\omega_{j,k}$ in Equation 7.

$$\begin{aligned} TE(d,t) = e^{i(\omega t)}, &\omega = \omega_{j,k} \cdot r_s^{-1}(d) \\ j = 0, ..., d_{TV} - 1, &k = 0, ..., K-1 \end{aligned} \tag{7}$$

In TE, $\omega_{j,k}$ means the time vector in dimension j has K frequencies. But in Equation 2 of TV, the frequency of time vector in dimension i is fixed with c_i.

The Relations between Different Mechanisms

Time encoding with different sampling rates is related to time vector with fixed sampling rate and a general complex expression [Wang *et al.*, 2020].

- TV is a special case of TE. If we set $\omega_{k,j} = c_i$, then $TE(d,t) = TV(t, 2i+1) + iTV(t, 2i)$.
- TE is a special case of a fundamental complex expression $r \cdot e^{i \cdot (\omega x + \theta)}$. We set $\theta = 0$ as we focus more on the relation between different time points than the value of the

first point; We understand term r as the representation of observations and leave it to learn by computing models. Besides, TE inherits the properties of position-free offset transformation and boundedness [Wang *et al.*, 2020].

3.3 Injecting Time Encoding mechanism into Echo State Network

Echo state network is a fast and efficient recurrent neural network. A typical ESN consists of an input layer $W_{in} \in R^{N\times D}$, a recurrent layer, called reservoir $W_{res} \in R^{N\times N}$, and an output layer $W_{out} \in R^{M\times N}$. The connection weights of the input layer and the reservoir layer are fixed after initialization, and the output weights are trainable. $u(t) \in R^D$, $x(t) \in R^N$ and $y(t) \in R^M$ denote the input value, reservoir state and output value at time t, respectively. The state transition equation is:

$$\begin{aligned} x(t) &= f(W_{in}u(t) + W_{res}x(t-1)) \\ y(t) &= W_{out}x(t) \end{aligned} \tag{8}$$

Before training, three are three main hyper-parameters of ESNs: Input scale w^{in}; Sparsity of reservoir weight α; Spectral radius of reservoir weight $\rho(W_{res})$ [Jiang and Lai, 2019].

However, existing ESNs-based methods cannot model the irregularities of ISTS. Thus, we make up for this by proposing Time Encoding Echo State Network (TE-ESN).

Time Encoding Phase

TE-ESN has D reservoirs, assigning each time series of input an independent reservoir. An observation u_t^d is transferred trough input weight W_{in}^d, time encoding $TE(d,t)$, reservoir weight W_{res}^d and output weight W_{out}^d. The structure of TE-ESN is shown in Figure 1. The state transition equation is:

$$\begin{aligned} \tilde{x}_t^d &= \gamma_f x_t^{d'} + (1-\gamma_f)x^{D\setminus d} && \textit{Reservoir} \\ x_t^{d'} &= \gamma_l x_t^d + (1-\gamma_l)(x_{t-1}^d + x_{t-k}^d) && \textit{Long short} \\ x_t^d &= \tanh(TE(d,t) + W_{in}^d u_t^d + W_{res}^d \tilde{x}_{t-1}^d) && \textit{Time encoding} \\ x^{D\setminus d} &= \frac{1}{D-1}\sum_{i\in D\setminus d} \tilde{x}^i && \textit{Neighbor} \end{aligned} \tag{9}$$

TE-ESN creates three highlights compared with other ESNs-based methods by changing the *Reservoir state*:

- Time encoding mechanism (TE). TE-ESN integrates time information when modeling the dynamic dependencies of input, by changing recurrent states in reservoirs through TE term to *Time encoding state*.
- Long short-term memory mechanism (LS). TE-ESN leans different temporal span dependencies, by incorporating not only short-term memories from state in last time, but also long-term memories from state in former k time (k is the time skip) to *Long short state*.
- Series fusion (SF). TE-ESN also considers the horizontal information between time series, by changing *Reservoir state* according to not only the former state in its time series but also the *Neighbor state* in other time series.

The coefficients γ_l and γ_f trade off the memory length in *Long short state* and the fusion intensity in *Reservoir state*.

Algorithm 1 TE-ESN

Input: $U_{train} = \{u_t^d, t\}$: training input;
$Y_{train} = y_t^d, t$: teacher signal;
$U_{test} = \{u_t^d, t\}$: test input;
MT: maximum time;
γ_l: leaky rate; γ_f: fusion rate;
k: long term time span;
λ: regularization coefficient;
w_{in}: input scale of W_{in};
$\rho(W_{res})$: spectral radius of W_{res};
α: sparsity of W_{res};
Output: Y_{pre}: prediction result.

1: Randomly initialized W_{in} in $[-w^{in}, w^{in}]$;
2: Randomly initialized W_{res} with α and $\rho(W_{res})$.
3: **for** $i = 1$ to $|U_{train}|$ **do**
4: **for** $t = 1$ to MT **do**
5: Compute $TE(d,t)$ by Equation 7
6: Compute $\tilde{x}(t)$ by Equation 9
7: **end for**
8: **end for**
9: $\tilde{X} = \{\tilde{x}(t)\}$
10: $TE = \{TE(t)\}$
11: Compute W_{out} by Equation 11
12: **for** $t = 1$ to T_{test} **do**
13: Compute $TE_{test}(d,t)$ by Equation 7
14: Compute $\tilde{x}_{test}(t)$ by Equation 9
15: **end for**
16: $\tilde{X}_{test} = \{\tilde{x}_{test}(t)\}$
17: $TE_{test} = \{TE_{test}(t)\}$
18: $Y_{pre} = W_{out}(\tilde{X}_{test} - TE_{test})$

Time Decoding Phase

The states in reservoir of TE-ESN have time information as TE embeds time codes into the representations of model input. For final value prediction, it should decode the time information and get the real estimated value at time t_{pre} by Equation 10. Further, by changing the time t_{pre}, we can get different prediction results in different time points.

$$y(t_{pre}) = W_{out}(\tilde{x}(t) - TE(t_{pre})) \tag{10}$$

Equation 11 is the calculation formula of the readout weights when training to find a solution to the least squares problem with regularization parameter λ.

$$\begin{aligned} &\min_{W_{out}} ||Y_{pre} - Y||_2^2 + \lambda||W_{out}||_2^2 \\ &W_{out} = Y(\tilde{X} - TE)^T((\tilde{X} - TE)(\tilde{X} - TE)^T + \lambda I)^{-1} \end{aligned} \tag{11}$$

Algorithm 1 shows the process of using TE-ESN for prediction. Line 1-11 obtains the solution of readout weights W_{out} of TE-ESN by that using the training data. Line 12-18 shows the way to predict the value of test data. Assuming the reservoir size of TE-ESN is fixed by N, The maximum time MT is T, the input has D time series, the complexity is:

$$C = O(\alpha TN^2 + TND) \tag{12}$$

4 Experiments

4.1 Datasets

- *MG* [Mackey and Glass, 1977] is a chaotic system. $y(t+1) = y(t) + \delta(a\frac{y(t-\frac{\tau}{\delta})}{1+y(t-\frac{\tau}{\delta})^n} - by(t))$. $\delta, a, b, n, \tau, y(0) =$

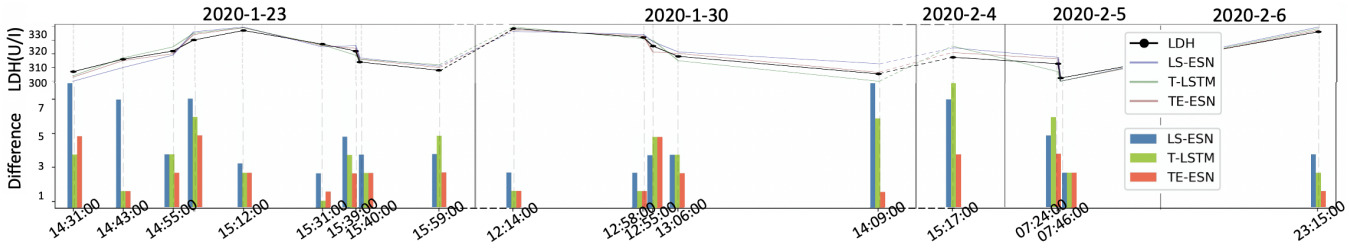

Figure 2: Lactic dehydrogenase (LDH) forecasting for a 70-year-old female COVID-19 patient

	BPRNNs-based			ESNs-based				Ours	
	M-RNN	T-LSTM	GRU-D	ESN	leaky-ESN	DeepESN	LS-ESN	TV-ESN	TE-ESN
MG	0.232±0.005	0.216±0.003	0.223±0.005	0.229±0.001	0.213±0.001	0.197±0.000	0.198±0.000	0.204±0.001	**0.195±0.001**
SILSO	2.950±0.740	2.930±0.810	2.990±0.690	3.070±0.630	2.950±0.590	2.800±0.730	2.540±0.690	2.540±0.790	**2.390±0.780**
USHCN	0.752±0.320	0.746±0.330	0.747±0.250	0.868±0.290	0.857±0.200	0.643±0.120	0.663±0.150	0.647±0.150	**0.640±0.190**
COVID-19	0.098±0.005	0.096±0.007	0.100±0.005	0.136±0.006	0.135±0.007	0.129±0.006	0.120±0.007	0.115±0.005	**0.093±0.005**
	0.959±0.004	0.963±0.003	0.963±0.004	0.941±0.003	0.942±0.003	0.948±0.003	0.949±0.003	0.958±0.002	**0.965±0.002**

Table 1: Prediction results of nine methods on four datasets (COVID-19 mortality in AUC-ROC; Others in MSE)

$0.1, 0.2, -0.1, 10, 17, 1.2$. t random increases with irregular interval. The task is one-step-ahead-forecasting.

- *SILSO* [Center, 2016] provides open-source monthly sunspot series from 1749 to 2020. It has irregular time intervals, from 1 to 6 month. The task is one-step-ahead forecasting from 1980 to 2019.
- *USHCN* [Menne and R., 2010] consists of daily meteorological data of 48 states from 1887 to 2014. Irregular time intervals are from 1 to 7 days. Sampling rates are from 0.33 to 1 per day. We use the records of 4 neighboring states to predict New York temperature in next 7 days.
- *COVID-19* [Yan L, 2020] contains patients' blood samples from 10 Jan. to 18 Feb. 2020 at Tongji Hospital, Wuhan, China. It has 80 features from 485 patients with 6877 records. Irregular time intervals are from 1 minus to 12 days. Sampling rates are from 0 to 6 per day. The task is to early predict in-hospital mortality before 24 hours and one-step-ahead forecasting for each biomarkers.

4.2 Baselines

- *BPRNNs-based*: There are 3 methods designed for ISTS data with BP training - M-RNN [Jinsung *et al.*, 2017], T-LSTM [Baytas *et al.*, 2017] and GRU-D [Che *et al.*, 2018]. Each of them have be introduced in Section 2.
- *ESNs-based*: There are 4 methods designed based on ESNs - ESN [Jaeger, 2002], Leaky-ESN [Jaeger *et al.*, 2007], DeepESN [Gallicchio *et al.*, 2017] and LS-ESN [Zheng *et al.*, 2020]. Each of them have be introduced in Section 2.
- *Our methods*: We use TV-ESN with the time representation embedded by TV, we use TE-ESN with the time representation embedded by TE.

4.3 Experiment Setting

We use Genetic Algorithms (GA) [Zhong *et al.*, 2017] to optimize hyper-parameters shown in Table 5. For TV-ESN, we set $\omega = c_i, d_{TV} = 64$. For TE-ESN, We set $\omega_{k,j} = M_j^{-\frac{2j}{d_{TE}}}$, where $M_0 = \frac{MT}{2}, M_1 = MT, M_2 = 2MT, M_3 = 4MT$ and $d_{TE} = 64$. Results are got by 5-fold cross validation. Method performances are evaluated by AUC-ROC and MSE. Network property of ESNs is evaluated by Memory Capability (MC) [Farkas *et al.*, 2016] in Equation 13. where r^2 is the squared correlation coefficient.

Parameters	Value range	Parameters	Value range
w^{in}, α, ρ	$(0, 1]$	γ_l, γ_f	$[0, 1]$
k	$\{2, 4, 6, 8, 10, 12\}$	λ	$\{10^{-4}, 10^{-2}, 1\}$

Table 2: Search settings of hyper-parameters

$$MC = \sum_{k=0}^{\infty} r^2(u(t-k), y(t)) \tag{13}$$

4.4 Results

The conclusions drawn from the results are shown in italics.

Prediction Results

Shown in Table 1: (1) TE-ESN outperforms all baselines on four datasets. It means *Learning two irregularities of ISTS helps for prediction and TE-ESN has this ability*. (2) TE-ESN is better than TV-ESN in multivariable time series datasets (COVID-19, USHCN) shows *the effect of Property 3 of TE*; TE-ESN is better than TV-ESN in univariable time series datasets (SILSO, MG) shows *the advantage of multiple frequencies options of TE*. (3) ESNs-based methods perform better in USHCN, SILSO and MG, while BPRNNs-based method performs better in COVID-19. Which shows *the characteristic of ESNs that they are good at modeling the consistent dynamic chaos system*, such as astronomical, meteorological and physical. Figure 2 shows a case of forecasting lactic dehydrogenase (LDH), an important bio-marker of

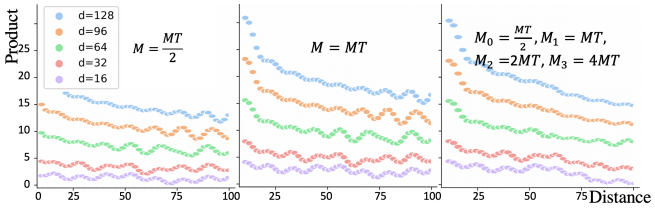

Figure 3: Dimension and frequency setting of time encoding

	$c_i, 32$	$c_i, 64$	$\omega_{d,i}, 32$	$\omega_{d,i}, 64$
MG	0.226±0.001	0.204±0.001	0.210±0.001	**0.193±0.001**
SILSO	2.690±0.600	2.540±0.79	2.550±0.750	**2.390±0.780**
USHCN	0.681±0.180	0.670±0.200	0.673±0.170	**0.640±0.190**
COVID-19	0.105±0.006	0.099±0.005	0.101±0.005	**0.093±0.005**
	0.949±0.002	0.952±0.003	0.950±0.002	**0.965±0.002**

Table 3: Prediction results of TE-ESN with different ω, d_{TE}

	w/o TE	w/o LS	w/o SF	TE-ESN
MG	0.210±0.001	0.213±0.001	0.193±0.001	**0.193±0.001**
SILSO	2.790±0.630	2.930±0.690	2.390±0.780	**2.390±0.780**
USHCN	0.713±0.120	0.757±0.210	0.693±0.160	**0.640±0.190**
COVID-19	0.135±0.006	0.130±0.006	0.125±0.007	**0.093±0.005**
	0.943±0.003	0.949±0.003	0.956±0.003	**0.965±0.002**

Table 4: Prediction results of TE-ESN with different mechanisms

COVID-19 [Yan L, 2020; Sun *et al.*, 2020c]. TE-ESN has smallest difference between real and predicted LDH values.

Time Encoding Mechanism Analysis

Dot product between two sinusoidal positional encoding decreases with increment of absolute value of distance [Yan *et al.*, 2019]. (1) Figure 3 shows the relation of TE dot product and time distance, it shows that *using multiple frequencies will enhance monotonous of negative correlation between dot product and distance*. (2) Table 3 shows the prediction results in different TE settings, results shows that *using multiple frequencies can improve the prediction accuracy*.

Ablation Study of TE-ESN

We test the effect of TE, LS and SF, which are introduced in Section 3.3, by removing TE term, setting $\gamma_l = 1$ and setting $\gamma_f = 1$. The results in Table 4 show that *all theses three mechanisms of TE-ESN contribute to the final prediction tasks*. TE has the greatest impact in COVID-19, the reason may be that the medical dataset has the strongest irregularity compared with other datasets. LS has the greatest impact in USHCN and SILSO, as there are many long time series, it is necessary to learn the dependence in different time spans. SF has a relatively small impact, the results have no change in SILSO and MG as they are univariate.

Hyper-Parameters Analysis of TE-ESN

In TE-ESN, each time series has a reservoir, reservoirs setting can be different. Figure 4 shows COVID-19 mortality prediction results when changing spectral radius ρ and time skip k of LDH and hs-CRP. *Setting uniform hyper-parameters or different hyper-parameters for each reservoir has little effect on the prediction results*. Thus, we set all reservoirs with the same hyper-parameters for efficiency. Table 5 shows the best hyper-parameter settings.

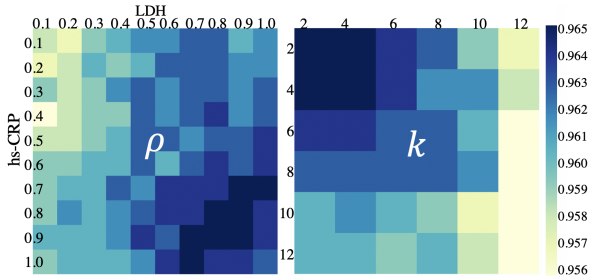

Figure 4: Mortality prediction of TS-ESN with different ρ and k

	w^{in}	α	ρ	γ_l	k	γ_f	λ
MG	1	0.1	0.7	0.8	6	1.0	10^{-2}
SILSO	1	0.1	0.6	0.8	10	1.0	10^{-2}
USHCN	1	0.1	0.7	0.8	12	0.8	10^{-2}
COVID-19	1	0.2	0.8	0.8	2	0.8	10^{-2}
	1	0.3	0.9	0.7	4	0.9	10^{-2}

Table 5: Best settings of hyper-parameters of TE-ESN

	ESN	leaky-ESN	DeepESN	LS-ESN	w/o TE	TE-ESN
MC	35.05	39.65	42.98	46.05	40.46	**47.83**

Table 6: Memory capacity results of ESNs-based methods

Memory Capability Analysis of TE-ESN

Memory capability (MC) can measure the short-term memory capacity of reservoir, an important property of ESNs [Gallicchio *et al.*, 2018]. Table 6 shows that TE-ESN obtains the best MC, and *TE mechanism can increase the memory capability*.

5 Conclusions

In this paper, we propose a novel Time Encoding (TE) mechanism in complex domain to model the time information of ISTS. It can represent the irregularities of intra-series and inter-series. We create a novel Time Encoding Echo State Network (TE-ESN), which is the first method to enable ESNs to handle ISTS. TE-ESN can model both longitudinal long short-term dependencies in time series and horizontal influences among time series. We evaluate the method and give several model related analysis in two prediction tasks on four datasets. The results show that TE-ESN outperforms the existing state-of-the-art and has good properties. Future works will focus on the dynamic reservoir properties and hyper-parameters optimization of TE-ESN, and will incorporate deep structures to TE-ESN for better prediction accuracy.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2020YFB2103402).

References

[Baytas *et al.*, 2017] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, and Jiayu Zhou. Patient subtyping via time-aware LSTM networks. In *KDD 2017*, pages 65–74, 2017.

[Center, 2016] SILSO World Data Center. The international sunspot number, int. sunspot number monthly bull. online catalogue (1749-2016). http://www.sidc.be/silso/, 2016.

[Che *et al.*, 2018] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Sentific Reports*, 8(1):6085, 2018.

[Farkas *et al.*, 2016] Igor Farkas, Radomír Bosák, and Peter Gergel. Computational analysis of memory capacity in echo state networks. *Neural Networks*, 83:109–120, 2016.

[Fawaz *et al.*, 2019] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Min. Knowl. Discov.*, 33(4):917–963, 2019.

[Gallicchio and Micheli, 2017] Claudio Gallicchio and Alessio Micheli. Deep echo state network (deepesn): A brief survey. *CoRR*, abs/1712.04323, 2017.

[Gallicchio *et al.*, 2017] Claudio Gallicchio, Alessio Micheli, and Luca Pedrelli. Deep reservoir computing: A critical experimental analysis. *Neurocomputing*, 268(dec.11):87–99, 2017.

[Gallicchio *et al.*, 2018] Claudio Gallicchio, Alessio Micheli, and Luca Pedrelli. Design of deep echo state networks. *Neural Networks*, 108:33–47, 2018.

[Gehring *et al.*, 2017] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In *ICML 2017*, pages 1243–1252, 2017.

[Hao and Cao, 2020] Yifan Hao and Huiping Cao. A new attention mechanism to classify multivariate time series. In *IJCAI 2020*, pages 1999–2005. ijcai.org, 2020.

[Jaeger *et al.*, 2007] Herbert Jaeger, Mantas Lukosevicius, Dan Popovici, and Udo Siewert. Optimization and applications of echo state networks with leaky- integrator neurons. *Neural Networks*, 20(3):335–352, 2007.

[Jaeger, 2002] Herbert Jaeger. Adaptive nonlinear system identification with echo state networks. In *NIPS 2002*, pages 593–600, 2002.

[Jiang and Lai, 2019] Jun-Jie Jiang and Ying-Cheng Lai. Model-free prediction of spatiotemporal dynamical systems with recurrent neural networks: Role of network spectral radius. *CoRR*, abs/1910.04426, 2019.

[Jinsung *et al.*, 2017] Yoon Jinsung, Zame William R., and Mihaelaå Van Der Schaar. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Trans. Biomed. Eng.*, PP:1–1, 2017.

[Karim *et al.*, 2019] Fazle Karim, Somshubra Majumdar, and Houshang Darabi. Insights into LSTM fully convolutional networks for time series classification. *IEEE Access*, 7:67718–67725, 2019.

[Mackey and Glass, 1977] M. Mackey and L Glass. Oscillation and chaos in physiological control systems. *Science*, 197(4300):287–289, 1977.

[Menne and R., 2010] Williams C. Menne, M. and Vose R. Long-term daily and monthly climate records from stations across the contiguous united states. https://cdiac.ess-dive.lbl.gov/epubs/ndp/ushcn/ushcn.html, 2010.

[Shukla and Marlin, 2019] Satya Narayan Shukla and Benjamin M. Marlin. Interpolation-prediction networks for irregularly sampled time series. In *ICLR*, 2019.

[Sun *et al.*, 2020a] Chenxi Sun, Shenda Hong, and Hongyan Li. A review of designs and applications of echo state networks. *CoRR*, abs/2012.02974, 2020.

[Sun *et al.*, 2020b] Chenxi Sun, Shenda Hong, Moxian Song, and Hongyan Li. A review of deep learning methods for irregularly sampled medical time series data. *CoRR*, abs/2010.12493, 2020.

[Sun *et al.*, 2020c] Chenxi Sun, Shenda Hong, Moxian Song, Hongyan Li, and Zhenjie Wang. Predicting covid-19 disease progression and patient outcomes based on temporal deep learning. *BMC Medical Informatics and Decision Making*, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS 2017*, pages 5998–6008, 2017.

[Virk, 2006] Imran S. Virk. Diurnal blood pressure pattern and risk of congestive heart failure. *Congestive Heart Failure*, 12(6):350–351, 2006.

[Wang *et al.*, 2019] Yishu Wang, Ye Yuan, Yuliang Ma, and Guoren Wang. Time-dependent graphs: Definitions, applications, and algorithms. *Data Sci. Eng.*, 4(4):352–366, 2019.

[Wang *et al.*, 2020] Benyou Wang, Donghao Zhao, Qiuchi Li, Peng Zhang, and Jakob Grue Simonsen. Encoding word order in complex embeddings. In *ICLR*, 2020.

[Xing *et al.*, 2010] Zhengzheng Xing, Jian Pei, and Eamonn J. Keogh. A brief survey on sequence classification. *SIGKDD Explorations*, 12(1):40–48, 2010.

[Yan *et al.*, 2019] Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. TENER: adapting transformer encoder for named entity recognition. *CoRR*, abs/1911.04474, 2019.

[Yan L, 2020] Goncalves J et al. Yan L, Zhang H T. An interpretable mortality prediction model for covid-19 patients. *Nature, Machine intelligence*, 2, 2020.

[Zheng *et al.*, 2020] Kaihong Zheng, Bin Qian, Sen Li, Yong Xiao, Wanqing Zhuang, and Qianli Ma. Long-short term echo state network for time series prediction. *IEEE Access*, 8:91961–91974, 2020.

[Zhong *et al.*, 2017] Shisheng Zhong, Xiaolong Xie, Lin Lin, and Fang Wang. Genetic algorithm optimized double-reservoir echo state network for multi-regime time series prediction. *Neurocomputing*, 238:191–204, 2017.