DAMA-Net: A Novel Predictive Model for Irregularly Asynchronously and Sparsely Sampled Multivariate Time Series

Zhen Wang [1] **Yang Zhang** [1] **Ai Jiang** [2] **Ji Zhang** [3] **Zhao Li** [4] **Jun Gao** [5] **Ke Li** [6] **Chenhao Lu** [1]

Abstract

Irregularly, asynchronously and sparsely sampled multivariate time series (IASS-MTS) data occur naturally in practical domains. They are characterized by sparse non-uniform time intervals between successive observations and different sampling rates amongst series. These properties pose substantial challenges to contemporary machine learning models for learning complicated intra-series and inter-series relations within and across IASS-MTS. To address these challenges, we present a time-aware Dual-Attention and Memory-Augmented Networks architecture (DAMA-Net). The proposed model aims at leveraging both time irregularity, multi-sampling rates and global temporal patterns information inherent in time series so as to learn more effective representations and improve prediction performance. We evaluate our model on two real-world datasets for IASS-MTS classification tasks. The results show that our model outperforms state-of-the-art methods in terms of classification performance. Moreover, we conduct the ablation study to demonstrate the contribution made by different mechanisms and modules in our model.

1. Introduction

We study the problem of classification of irregularly, asynchronously and sparsely sampled sequences in this work. As illustrated in Fig.1, IASS-MTS is a sequence of samples characterized by 1) varying length time series records of observations (see Fig.1-a); 2) asynchronously sampled features within each time point (see Fig.1-b); 3) time sparsity when the intervals between observation times are large (see Fig.1-c). Such time series data arise in a number of scientific and industrial domains, including climate science (Shi et al., 2015), ecology (Clark & Bjørnstad, 2004), astronomy (Vio et al., 2000), finance (Eng & Gustafsson, 2007) and medicine (Che et al., 2018), where the observation process is constrained to a degree that prohibits regular observation/sampling. For example, in air pollution forecasting, the collected MTS data are usually 1) incomplete due to broken sensors, failed data transmissions or damaged storage; 2) asynchronous when the data is collected from channels having different time scales. These characteristics of IASS-MTS data pose multiple challenging threats to classical machine learning models and algorithms that require data to be defined in a coherent fixed-dimensional feature space with constant intervals between consecutive time steps.

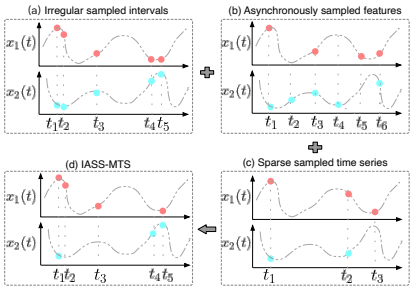

Figure 1. Irregularly asynchronously and sparsely sampled MTS.

There has been some progress on this problem. For example, (Neil et al., 2016) proposed Phased-LSTM to improve current RNN models, which extends the LSTM unit by adding a new time gate k_t to deal with irregularly sampled time series. (Che et al., 2018) proposed the GRU-D to incorporate irregular time intervals and handle asynchronously sampled features problem from a missing data perspective. However, it might fail to model very sparse samples because the model is only updated by samples lying in the model's active state during training. (Shukla & Marlin, 2019) presented the interpolation-prediction network (IPNet) framework. IP-Net applies several semi-parametric interpolation schemes to obtain a regularly sampled time series representation by multiple interpolants. (Horn et al., 2020) employed a set function-based approach for the task of classification on irregularly sampled and unaligned time series observations. Besides, other efforts have also been dedicated to this problem for dealing with such irregularity (Futoma et al., 2017; Yoon et al., 2018; Li & Marlin, 2020; Tan et al., 2020;

[1]Zhejiang Lab, China [2]University of St. Andrews, United Kingdom [3]University of Southern Queensland, Australia [4]Alibaba Group, China [5]Peking University, China [6]Dalian Maritime University. Correspondence to: Ji Zhang <zhangji77@gmail.com>.

Time Series Workshop the 38^{th} International Conference on Machine Learning, 2021.

Shukla & Marlin, 2021; Bianchi et al., 2019; Fortuin et al., 2020; Guo et al., 2019; Kidger et al., 2020; Xu et al., 2019; Soleimani et al., 2017).

Despite the improvement that existing approaches have achieved, some of the limitations still exist: 1) IASS-MTS measurements are frequently correlated both within series and across series. The current attention is mostly paid to modeling the intra-series dependency, whereas the interactions across series are not well studied. It has been noted that different measurements are often intertwined and this inter-series relationship is usually informative. For example, the blood pressure of a patient at a given time not only could be correlated with the blood pressure at other times, but it also could have a relation with the heart rate at that or other times; 2) most methods have other requirements which may not be satisfied in real IASS-MTS data. For example, many of them work on data with low missing rates, which suffers the failure when the missing ratio raises up or consecutive missing values occur (i.e., data is highly sparse).

To address the aforementioned limitations and challenges, we propose a novel DAMA-Net architecture to deal with the problem of asynchronous interactions, irregularity and sparsity of sampling intervals of IASS-MTS data. Specifically,

- We introduce a series-dependent intra-attention embedding module associated with a learned time encoding. It takes IASS-MTS data as the input and produces a fixed-length latent representation over a set of interpolants which encapsulate the intra-series interaction and circumvent asynchronously sampled features;
- Then, enhanced by modality indicator and position embedder within series, we build an inter-series attention module on the top of intra-attention embedding networks to effectively handle the interactions among different series across distinct time steps;
- Finally, we employ the external memory module for DAMA-Net to capture global temporal dynamics. The memory here can be interpreted as a container of highly summarized global structure information of sequence data. The DAMA-Net utilizes the knowledge of temporal patterns to construct global representations so as to mitigate the sparsity inherent in real-world IASS-MTS data and improve the prediction performance;
- We evaluate the proposed DAMA-Net model by conducting detailed comparative experiments and the ablation study on real-world datasets, which demonstrates the good performance of our proposed model.

To the best of our knowledge, this is the first work to propose time-aware dual attention and memory networks that jointly models the correlations of both within series and across series as well as global temporal dynamics for simultaneously handling the sparsity, asynchronicity and irregularity of sampling of multivariate time series.

2. Proposed Method

2.1. Problem Formulation

We let $\mathcal{D} = \{(\boldsymbol{s}_n, y_n) \in (\mathcal{S}, \mathcal{Y}) \mid n = 0, ..., N-1\}$ represent N data cases. Each data case consists of a D-dimensional irregularly, asynchronously and sparsely sampled multivariate time series $\boldsymbol{s}_n = \{\boldsymbol{s}_{n,d} \,|\, d = 1, ..., D\}$ as well as its label y_n. Each dimension of $\boldsymbol{s}_n$ is a univariate time series $\boldsymbol{s}_{n,d}$ (a.k.a a variable). We denote $T_{n,d}$ as the number of records of the d^{th} univariate time series of the n^{th} data case. Each univariate time series can be represented as a list of observed tuples $\boldsymbol{s}_{n,d} = [(t_{n,d,1}, x_{n,d,1}), (t_{n,d,2}, x_{n,d,2}), ..., (t_{n,d,T_{n,d}}, x_{n,d,T_{n,d}})]$, where $x_{n,d,T_{n,d}}$ is the observed value of the d^{th} variable at time step $T_{n,d}$ for multivariate time series data case n and $t_{n,d,T_{n,d}} \in \mathbb{R}_0^+$ is the corresponding observed time. We define $\boldsymbol{t}_{n,d} = [t_{n,d,1}, t_{n,d,2}, ..., t_{n,d,T_{n,d}}]$ to be the list of timestamps and $\boldsymbol{x}_{n,d} = [x_{n,d,1}, ..., x_{n,d,T_{n,d}}]$ to be the list of observations for the d^{th} univariate time series of data case n. For IASS-MTS data, different variables of the multivariate time series can have observations at different times, as well as different numbers of observations, which means $\boldsymbol{t}_{n,d} \neq \boldsymbol{t}_{n,d'}$ and $T_{n,d} \neq T_{n,d'}$ for $d \neq d'$ in general. We aim at learning a function $\mathscr{F} : \mathcal{S} \rightarrow \mathcal{Y}$ that can predict the target y_n given the multivariate time series $\boldsymbol{s}_n$.

2.2. Model Architecture

In this section, we elaborate on the proposed DAMA-Net model. Figure 2 shows its detailed architecture.

2.2.1. Time Encoding

To incorporate temporal irregularity information and improve the expressiveness of irregularly sampled time series data, inspired by (Kazemi et al., 2019), we learn a meaningful vector representation for continuous time τ. Specifically, the time encoding module transforms the 1-d time axis to a vector of size $k+1$ by

$$\Psi(\tau)[i] = \begin{cases} \omega_0 \tau + \varphi_0, & if\ i = 0 \\ sin(\omega_i \tau + \varphi_i), & if\ 1 \leq i \leq k \end{cases} \tag{1}$$

where $\Psi(\tau)[i]$ is the i^{th} dimension of embedding vector $\Psi(\tau)$, and $\{\omega_i, \varphi_i\}_{i=0}^{k}$ are learnable parameters. A sine function term helps capture periodic patterns while the linear term represents the progression of time and captures non-periodic behaviors in the time series. We can implement the time encoding component by employing a fully connected layer followed by a sinusoidal activation function.

2.2.2. Dimension-dependent Intra-Attention

The goal of intra-series attention module is to utilize the learned time embeddings to provide a collection of interpolants, defined at the L reference time points $\boldsymbol{\tau} = [\tau_1, ..., \tau_L]$, of each of the D univariate time series of ir-

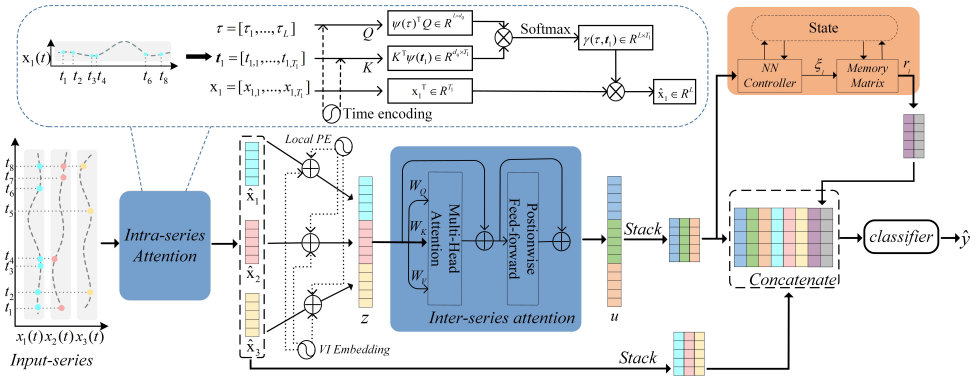

Figure 2. The overall architecture of the proposed DAMA-Net. DAMA-Net first learns continuous time point embeddings for preserving the informative varying intervals. Then DAMA-Net uses an intra-attention mechanism to accommodate irregular time observations. Next, the inter-series interactions across distinct time steps and modalities are captured by the proposed inter-attention module. In addition, an external memory component is designed to capture global temporal dynamics and store common knowledge extracted from the entire data, which will augment DAMA-Net architecture when time series are quite sparse. Finally, the outputs of the intra-attention module, inter-attention module and external memory module are concatenated together as the input of the classifier.

regularly, asynchronously and sparsely sampled multivariate time series. Our intra-series attention module separately transforms each dimension of multivariate time series. Each transformation is based on a time-attention mechanism which takes a query reference time point τ_l, a set of observed time point keys $\boldsymbol{t}_{n,d}$ and observed values $\boldsymbol{x}_{n,d}$ as input, and outputs an embedding $\hat{x}_{n,d,\tau_l}$ at time τ_l. Then, the interpolated multivariate time series is given by $\hat{\boldsymbol{s}}_n = \{\hat{\boldsymbol{s}}_{n,d} \mid d = 1, ..., D\}$, where $\hat{\boldsymbol{s}}_{n,d} = [(\tau_1, \hat{x}_{n,d,\tau_1}), ..., (\tau_L, \hat{x}_{n,d,\tau_L})]$.

$$\hat{x}_{n,d,\tau_l} = \sum_{j=1}^{T_{n,d}} \gamma(\tau_l, t_{n,d,j}) x_{n,d,j} \quad l = 1, ..., L \tag{2}$$

$$\gamma(\tau_l, t_{n,d,j}) = \frac{\exp\big(\psi(\tau_l)^T Q_d K_d^T \psi(t_{n,d,j})/\sqrt{d_q}\big)}{\sum_{j'=1}^{T_{n,d}} \exp\big(\psi(\tau_l)^T Q_d K_d^T \psi(t_{n,d,j'})/\sqrt{d_q}\big)} \tag{3}$$

In Eq.2, the interpolation weight $\gamma(\tau_l, t_{n,d,j})$ for each observation value is the inner product of query and keys with a softmax normalization, defined in Eq. 3. The dimension of query matrix Q_d and key matrix K_d are each of $(k+1)\times d_q$, and $1/\sqrt{d_q}$ is the scaling factor. In summary, intra-series attention embedding module naturally accommodates continuous time observations, encapsulates the intra-series interaction and provides a fixed-length latent representation for irregularly and asynchronously sampled MTS.

2.2.3. Indicator Enhanced Inter-Attention

The intra-attention module above ignores cross-series temporal correlations. Thus we proposed the inter-series attention in this section. As shown in Figure 2, the inter-attention module takes the chained sequence, $[\hat{\boldsymbol{x}}_{n,1}, ..., \hat{\boldsymbol{x}}_{n,D}]$, of all interpolated univariate time series as inputs to learn the cross-series relationships. To distinguish multiple univariate time series and incorporate the order information of interpolants, in this work, we 1) append a modality indicator d (MI) as an extra factor and learn an embedding $\mathcal{E}_d$ for it; 2) construct a *local* position embedding (PE_d^l), which indicates the position (l) of the interpolant within each separate univariate time series $\hat{\boldsymbol{s}}_{n,d}$.

$$\begin{aligned} PE_d(l, 2m) &= sin(l/10000^{2m/d_p}) \quad d = 1, ..., D \\ PE_d(l, 2m+1) &= cos(l/10000^{2m/d_p}) \quad d = 1, ..., D \end{aligned} \tag{4}$$

where l is the local position from 1 to L and $m = 0, ..., \lfloor \frac{d_p}{2} \rfloor$. Then the final input ($\boldsymbol{z}_n$) of attention module is calculated by summarizing the interpolant, modality indicator embedding and *local* positional embedding together as

$$\begin{aligned} \boldsymbol{z}_n = [\boldsymbol{z}_{n,1}, ..., \boldsymbol{z}_{n,D}], \quad \boldsymbol{z}_{n,d} = [z_{n,d,\tau_1}, ..., z_{n,d,\tau_L}] \\ z_{n,d,\tau_l} = \hat{x}_{n,d,\tau_l} + \mathcal{E}_d + PE_d^l \end{aligned} \tag{5}$$

And the inter-series attention function (IAF) is presented as

$$IAF(\boldsymbol{z}_n, W_Q, W_K, W_V) = \text{softmax}\Big(\frac{\boldsymbol{z}_n^T W_Q W_K^T \boldsymbol{z}_n}{\sqrt{d_p}}\Big) \boldsymbol{z}_n^T W_V \tag{6}$$

Definitely, we can use the multi-head attention here. Based on the inter-series attention block, we employ a position-wise feed-forward network and residual connections to construct a complete inter-series attention module. We denote $\boldsymbol{u}_n = [\boldsymbol{u}_{n,:,\tau_1}, ..., \boldsymbol{u}_{n,:,\tau_L}] \in \mathbb{R}^{D\times L}$ as the stacked output of inter-series attention module, where $\boldsymbol{u}_{n,:,\tau_l} = [u_{n,1,\tau_l}, ..., u_{n,D,\tau_l}]^T$. The main purpose of proposed module is to capture relationships among different univariate time series, thus boosting module's representation ability.

2.2.4. External Dynamic Memory

The basic idea of the memory module is to learn a parameterized memory matrix which caches global temporal knowledge. We use the similar approach as in DNC (Graves et al., 2016). Given the multivariate sequence $\boldsymbol{u}_n$ from the inter-series attention module, at the current time step l, we concatenate the vector $\boldsymbol{u}_{n,:,l}$ and a set of memory read vectors $\{\boldsymbol{r}_{n,1}, ..., \boldsymbol{r}_{n,l-1}\}$ from previous time steps as the input, then feed it into a neural network controller to obtain the interface vector $\boldsymbol{\xi}_{n,l}$. The interface vector $\boldsymbol{\xi}_{n,l}$ is split into

Table 1. Classification Performance.

MODEL	PHYSIONET	HUMAN ACTIVITY
PHASED-LSTM	0.836± 0.003	0.855± 0.005
GRU-IMPUTATION	0.764± 0.016	0.859± 0.004
GRU-δ_t	0.787± 0.014	0.857± 0.002
GRU-DECAY	0.807± 0.003	0.860± 0.005
GRU-D	0.818± 0.008	0.862± 0.005
IPNET	0.819± 0.006	0.869± 0.007
SEFT	0.795± 0.015	0.815± 0.002
MTAND	0.858± 0.004	0.907± 0.002
ODE-RNN	0.833± 0.009	0.885± 0.008
L-ODE-ODE ENC.	0.829± 0.004	0.870± 0.028
L-ODE-RNN ENC.	0.781± 0.018	0.838± 0.004
DAMA-NET	**0.871± 0.007**	**0.915± 0.004**

the interface parameters which are used to produce write vector $\boldsymbol{v}_{n,l}$, write weighting $w^w_{n,l}$, erase vector $\boldsymbol{e}_{n,l}$ and read weighting $w^r_{n,l}$ for updating memory $\boldsymbol{\mathcal{M}}_l$ by

$$\boldsymbol{\mathcal{M}}_l = \boldsymbol{\mathcal{M}}_{l-1} \circ (E - w^w_{n,l} e^T_{n,l}) + w^w_{n,l} v^T_{n,l} \tag{7}$$

and returning the l^{th} read vector $\boldsymbol{r}_{n,l}$ by

$$r_{n,l} = \boldsymbol{\mathcal{M}}^T_l w^r_{n,l} \tag{8}$$

We let $\boldsymbol{r}_n = [\boldsymbol{r}_{n,1}, ..., \boldsymbol{r}_{n,L}] \in \mathbb{R}^{A \times L}$ represent the final concatenated L read vectors from all time steps. In summary, we use an external memory module to capture desired global temporal dynamics extracted from the whole dataset so as to alleviate the high sparsity of sampling intervals.

2.2.5. PREDICTION NETWORK AND LEARNING

The intra-series attention embeddings $\hat{\boldsymbol{x}}_n$, read vectors $\boldsymbol{r}_n$ and inter-series attention module outputs $\boldsymbol{u}_n$ are concatenated together as $E_n = Concat(\hat{\boldsymbol{x}}_n, \boldsymbol{r}_n, \boldsymbol{u}_n) \in \mathbb{R}^{(2D+A)\times L}$, which are fed into the final classifier to make the prediction $\hat{y}_n = g_\theta(E_n)$. Our entire model is trained by minimizing the following focal loss function $\mathcal{L} =$

$$-\sum_{n=1}^{N}\sum_{c=1}^{C}[\alpha y^c_n (1-\hat{y}^c_n)^\beta \log \hat{y}^c_n + (1-\alpha)(1-y^c_n)\hat{y}^{c\beta}_n \log(1-\hat{y}^c_n)] \tag{9}$$

where C is the number of classes, true label $y^c_n \in \{0, 1\}$ and $\hat{y}^c_n$ is the estimated probability over the c^{th} class. α is the balancing weighting factor and β is the focusing parameter.

3. Experiments

We conduct experiments on two benchmark datasets (Shukla & Marlin, 2020): PhysioNet challenge 2012 dataset and Human Activity dataset. We compare our model with 11 representative baselines: Phased-LSTM (Neil et al., 2016), GRU-Imputation, GRU-δ_t, GRU-Decay, GRU-D (Che et al., 2018), L-ODE-RNN enc. (Chen et al., 2018), IPNet (Shukla & Marlin, 2019), SeFT (Horn et al., 2020), mTAND (Shukla & Marlin, 2021), ODE-RNN, L-ODE-ODE enc. (Rubanova et al., 2019). The dataset description, experimental setup and implementation details are given in the Appendix.

Table 2. Ablation Study.

DAMA-NET	PHYSIONET	HUMAN ACTIVITY
W/O INTRA-ATTN	0.8557± 0.0080	0.9009± 0.0051
W/O INTER-ATTN	0.8673± 0.0029	0.9100± 0.0043
W/O EXT. MEM.	0.8691± 0.0063	0.9130± 0.0019
W/O LOCAL PE	0.8701± 0.0055	0.9129± 0.0028
W/O MI	0.8682± 0.0074	0.9145± 0.0025

Results Table 1 summarizes the predictive results of all the models on PhysioNet mortality and Human activity classification task. The results show that our DAMA-Net model yields the best classification performance on both datasets (AUC score for imbalanced PhysioNet and accuracy rate for Human Activity), compared with various competitors. Take results in PhysioNet for example, our DAMA-Net model outperforms the RNN-Impute model by approximately 11% of AUC score, and improves the current best result by 1.3%. Those competitive experiment results indicate the effectiveness of our proposed model on irregularly asynchronously sparsely sampled multivariate time series classification tasks.

Ablation Study To further study the influence of the individual components in DAMA-Net, we perform an ablation study on our model by removing a specific module once at a time. The results are shown in Table 2. It can be seen that: 1) all the three modules (intra-series attention, inter-series attention and dynamic memory) in our model are necessary. We design a dual-attention mechanism to model the temporal information and capture the complex interactions within and across time series in the irregular and asynchronous setting to learn good representations of IASS-MTS data. Also, modeling global knowledge with an external memory module benefits the classification task; 2) In addition, introducing other techniques such as local positional embedding and modality embedding does indeed help improve the model's performance. This validates the importance of position-wise information and modality indicator.

4. Conclusion

In this paper, we propose a novel DAMA-Net model for classifying time series with asynchronous sampling, time irregularity and sparsity. Specifically, we introduce an intra-series attention interpolation module with a learned time encoding for capturing temporal intra-series interactions and dealing with asynchronism. A modality indicator enhanced inter-series attention module is used to learn interactions among different series across distinct time steps. Further, we employ the external memory module to represent common dynamics from the data for partially alleviating the data sparsity problem. Our approach yields the best performance on real-world datasets compared to various competitors.

References

Bianchi, F. M., Livi, L., Mikalsen, K. Ø., Kampffmeyer, M., and Jenssen, R. Learning representations of multivariate time series with missing data. *Pattern Recognition*, 96: 106973, 2019.

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. Neural ordinary differential equations. *Advances in neural information processing systems*, pp. 6571–6583, 2018.

Clark, J. S. and Bjørnstad, O. N. Population time series: process variability, observation errors, missing values, lags, and hidden states. *Ecology*, 85(11):3140–3150, 2004.

Eng, F. and Gustafsson, F. Algorithms for downsampling non-uniformly sampled data. In *2007 15th European Signal Processing Conference*, pp. 1965–1969. IEEE, 2007.

Fortuin, V., Baranchuk, D., Rätsch, G., and Mandt, S. Gp-vae: Deep probabilistic time series imputation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1651–1661. PMLR, 2020.

Futoma, J., Hariharan, S., and Heller, K. Learning to detect sepsis with a multitask gaussian process rnn classifier. In *International Conference on Machine Learning*, pp. 1174–1182. PMLR, 2017.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

Guo, Z., Wan, Y., and Ye, H. A data imputation method for multivariate time series based on generative adversarial network. *Neurocomputing*, 360:185–197, 2019.

Horn, M., Moor, M., Bock, C., Rieck, B., and Borgwardt, K. Set functions for time series. In *International Conference on Machine Learning*, pp. 4353–4363. PMLR, 2020.

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.

Kidger, P., Morrill, J., Foster, J., and Lyons, T. Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems*, 2020.

Li, S. C.-X. and Marlin, B. Learning from irregularly-sampled time series: A missing data perspective. In *International Conference on Machine Learning*, pp. 5937–5946. PMLR, 2020.

Neil, D., Pfeiffer, M., and Liu, S.-C. Phased lstm: Accelerating recurrent network training for long or event-based sequences. *arXiv preprint arXiv:1610.09513*, 2016.

Rubanova, Y., Chen, R. T., and Duvenaud, D. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, pp. 5320–5330, 2019.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

Shukla, S. N. and Marlin, B. M. Interpolation-prediction networks for irregularly sampled time series. *arXiv preprint arXiv:1909.07782*, 2019.

Shukla, S. N. and Marlin, B. M. A survey on principles, models and methods for learning from irregularly sampled time series: From discretization to attention and invariance. *arXiv preprint arXiv:2012.00168*, 2020.

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

Soleimani, H., Hensman, J., and Saria, S. Scalable joint models for reliable uncertainty-aware event prediction. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1948–1963, 2017.

Tan, Q., Ye, M., Yang, B., Liu, S., Ma, A. J., Yip, T. C.-F., Wong, G. L.-H., and Yuen, P. Data-gru: Dual-attention time-aware gated recurrent unit for irregular multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 930–937, 2020.

Vio, R., Strohmer, T., and Wamsteker, W. On the reconstruction of irregularly sampled time series. *Publications of the Astronomical Society of the Pacific*, 112(767):74, 2000.

Xu, D., Ruan, C., Kumar, S., Korpeoglu, E., and Achan, K. Self-attention with functional time representation learning. *Advances in neural information processing systems*, 2019.

Yoon, J., Zame, W. R., and van der Schaar, M. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 66(5):1477–1490, 2018.

A. Appendix

A.1. Datasets and Competitors

PhysioNet The PhysioNet challenge 2012 data set[1] consists of multivariate clinical time series data with 36 temporal variables (e.g. *Albumin*, *Glucose*, *pH*) and 6 general patient descriptors (e.g. *Age*, *RecordID*, *ICUType*) extracted from 12,000 intensive care unit (ICU) records. Set A comprised of four thousand records with outcome-related descriptors are available to challenge participants. Each record contains irregularly, asynchronously and sparsely sampled multivariate time series during the first 48 hours after admission to the ICU. The missing rate is approximately 80.5%, and the mortality labels are imbalanced at an approximate ratio of $1:6$ $(pos:neg)$. We trained and evaluated our method and competing methods on the binary mortality prediction task of whether a patient will die during the hospital stay. We evaluate model performance by using area under the ROC curve (AUC score) for imbalanced classification problems.

Human Activity Human Activity data set[2] contains 3D positions of the belt, chest and ankles (12 features in total) collected from five individuals performing various activities: walking, sitting, lying, standing, etc. We follow the data preprocessing procedure in (Rubanova et al., 2019), and construct a data set of 6,554 sequences with 12 channels and 50 time points. Labels are provided for each observation time point and denote the type of activity that the person is performing, such as walking, sitting, lying, etc. (11 classes in total). The task is to classify each time point by the type of activity. We report multi-class classification accuracy for evaluating the model performance.

Competitors The methods in our comparative evaluation are listed as follows:

Phased-LSTM: extending the LSTM unit by adding a new time gate k_t to process inputs sampled at asynchronous times (Neil et al., 2016).

GRU-Imputation: replacing each missing observation with weighted average of the mean of the variable across the training data and the last measurement within that variable (Che et al., 2018).

GRU-δ_t: concatenating the measurement, masking variable m_t and time interval δ_t (indicating how long the particular observation has been missing) as the input of the model (Che et al., 2018).

GRU-Decay: decaying the previous hidden state $h_{t\text{-}1}$ with a factor γ_h before using it to compute the new hidden state h_t (Che et al., 2018).

GRU-D: using both input decay mechanism and RNN hidden state decay mechanism (Che et al., 2018).

IPNet: employing a semi-parametric RBF interpolation network followed by the application of a prediction network (Shukla & Marlin, 2019).

ODE-RNN: using neural ordinary differential equations to specify hidden state dynamics, and updating the hidden state using a standard RNN (Rubanova et al., 2019).

L-ODE-ODE enc.: defining a latent ODE model using an ODE-RNN as the encoder and neural ODE as the decoder (Rubanova et al., 2019).

L-ODE-RNN enc.: proposing a latent-variable time series model using a variational autoencoder framework (Chen et al., 2018).

SeFT: representing the irregularly sampled time series data based on differentiable set function learning (Horn et al., 2020).

mTAND: presenting a multi-time attention module followed by a VAE-based encoder-decoder model for learning from sparse and irregularly sampled data (Shukla & Marlin, 2021).

A.2. Experimental Setup

In our DAMA-Net model, for PhysioNet dataset, we use a one-layer of $1D$ convolutional layer to process the concatenated sequence, and apply a fully connected network with soft-max regressor on the top of convolution output to do classification. For human activity dataset, we adopt a one-layer GRU to model the concatenated sequence, and also use a one-layer fully connected network with the soft-max regressor on each hidden state to produce the classification at each time-point. For a fair comparison, we use same experimental protocols as (Shukla & Marlin, 2021) and compare our experimental results with theirs. For each data set, 80% instances are randomly selected as the training set, and the remaining 20% are used for testing set. We use 20% of the training data for validation. We repeat each experiment five times and report their average performance by mean and standard deviation.

A.3. Implementation Details

In our DAMA-Net model, the hyperparameters are tuned on the validation set by grid search. The α and β of focal loss are 0.25 and 2 respectively. We learn time embedding size of $k+1 = 128$ and choose $L = 128$ reference time points. The dimension of intra-series attention query matrix Q and key matrix K are each $128 * 128$. The GRU hidden state size of the classifier of Human activity prediction task is 512. The $1D$ convolution kernel size and $1D$ Max pooling operation size of PhysioNet classifier are 3 and 2 respectively. The size of dynamic external memory matrix is $5 * 10$ with 2

[1] https://physionet.org/content/challenge-2012/

[2] https://archive.ics.uci.edu/ml/datasets.php

Algorithm 1 Training DAMA-Net on Human Activity dataset

Input: data $\mathcal{D} = \{(\boldsymbol{s}_n, y_n) \in (\mathcal{S}, \mathcal{Y}) \mid n = 0, ..., N-1\}$, α, β, lr, k, L, $nItrs$, $nHiddens$, nr_cells, $cell\_size$, $read\_heads$
Output: $Accuracy\ Rate$
procedure TRAINING()
  Initialize ω_i, φ_i, Q, K, W_Q, W_K, W_V, $\boldsymbol{\mathcal{M}}_0$
  for $itr = 1$ **to** $nItrs$ **do**
    for $(\boldsymbol{s}_n, y_n)$ in training set **do**
      $\{\boldsymbol{x}_{n,d}, \boldsymbol{t}_{n,d}\}_{d=1}^{D} \leftarrow \boldsymbol{s}_n$
      $\{\psi(\tau_l)\}_{l=1}^{L} \leftarrow$ time encoding of $\boldsymbol{\tau} = [\tau_1, ..., \tau_M]$
      $\{\psi(t_{n,d,j})\}_{d,j=1,1}^{D,T_{n,d}} \leftarrow$ time encoding of $\boldsymbol{t}_n$
      $\{\gamma(\tau_l, t_{n,d,j}), \hat{x}_{n,d,\tau_l}\}_{d,l=1,1}^{D,L} \leftarrow$ Eqns. 3 and 2
      $PE_d^l \leftarrow$ local positional embedding of $\hat{x}_{n,d,\tau_l}$
      $\mathcal{E}_d \leftarrow$ modality indicator embedding of d
      $\{z_{n,d,\tau_l}\}_{d,l=1,1}^{D,L} \leftarrow$ Eqn. 5
      $\boldsymbol{u}_n \leftarrow IAF(\boldsymbol{z}_n, W_Q, W_K, W_V)$
      $\boldsymbol{\xi}_{n,l} \leftarrow \hat{\boldsymbol{x}}_{n,:,\tau_m}, \{\boldsymbol{r}_{n,1}, ..., \boldsymbol{r}_{n,l-1}\}$
      $\boldsymbol{v}_{n,l}, w_{n,l}^{w}, \boldsymbol{e}_{n,l}, w_{n,l}^{r} \leftarrow \boldsymbol{\xi}_{n,l}$
      update memory $\boldsymbol{\mathcal{M}}_l$ by Eqn. 7
      yield read vector $r_{n,l} \leftarrow$ Eqn. 8
      $E_n \leftarrow Concat(\hat{\boldsymbol{x}}_n, \boldsymbol{r}_n, \boldsymbol{u}_n)$
      $\hat{y}_n \leftarrow$ classifier $g_\theta(E_n)$
      compute loss by Eqn. 9
      update DAMA-Net model parameters using BP
    end for
    compute accuracy rate (Acc) on validation set
  end for
  select best model by $\arg\max_{model}\{Acc\}$
end procedure
procedure TESTING()
  compute Acc on test set under selected best model
  return Acc
end procedure

read heads. The model is trained using the Adam optimizer with a learning rate of 0.0001. We implemented our model in Pytorch.

A.4. Algorithms

The pseudo-code of our model on the Human Activity data set is summarized as Algorithm 1.