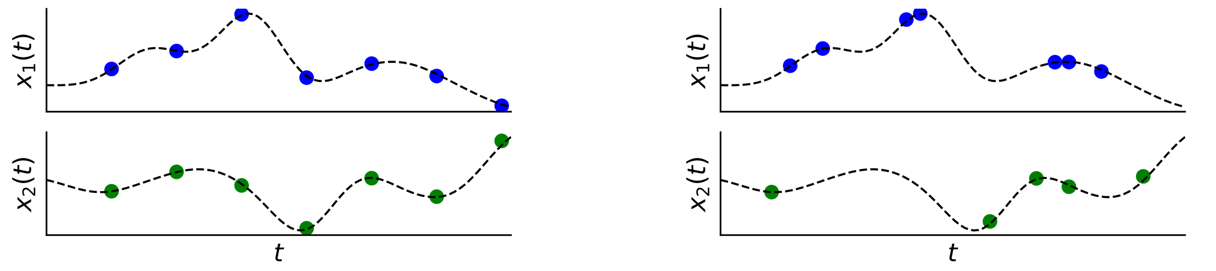
Multi-Time Attention Networks for Irregularly Sampled Time Series

Satya Narayan Shukla and Benjamin M. Marlin

{snshukla,marlin}@cs.umass.edu

Introduction

Problem: Learning from sparse and irregularly sampled multivariate time series.



Multivariate regularly (left) and irregularly (right) sampled time series

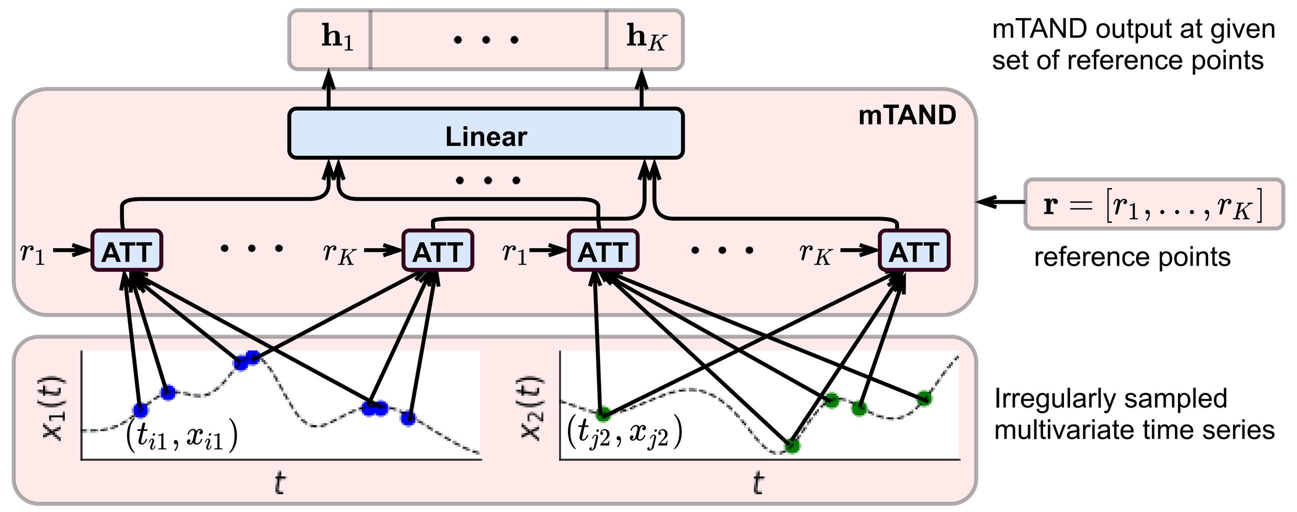
Challenges:

- Irregular spacing between observation time points
- Variable numbers of observations
- Lack of alignment of observation time points

Contributions:

- Proposed a flexible approach to modeling sparse and irregularly sampled time series data by leveraging a time attention mechanism.
- Temporally distributed latent representation to better capture local structure in time series data.
- Improved interpolation and classification performance current SOTA methods, while providing significantly reduced training times.

Multi-Time Attention Network

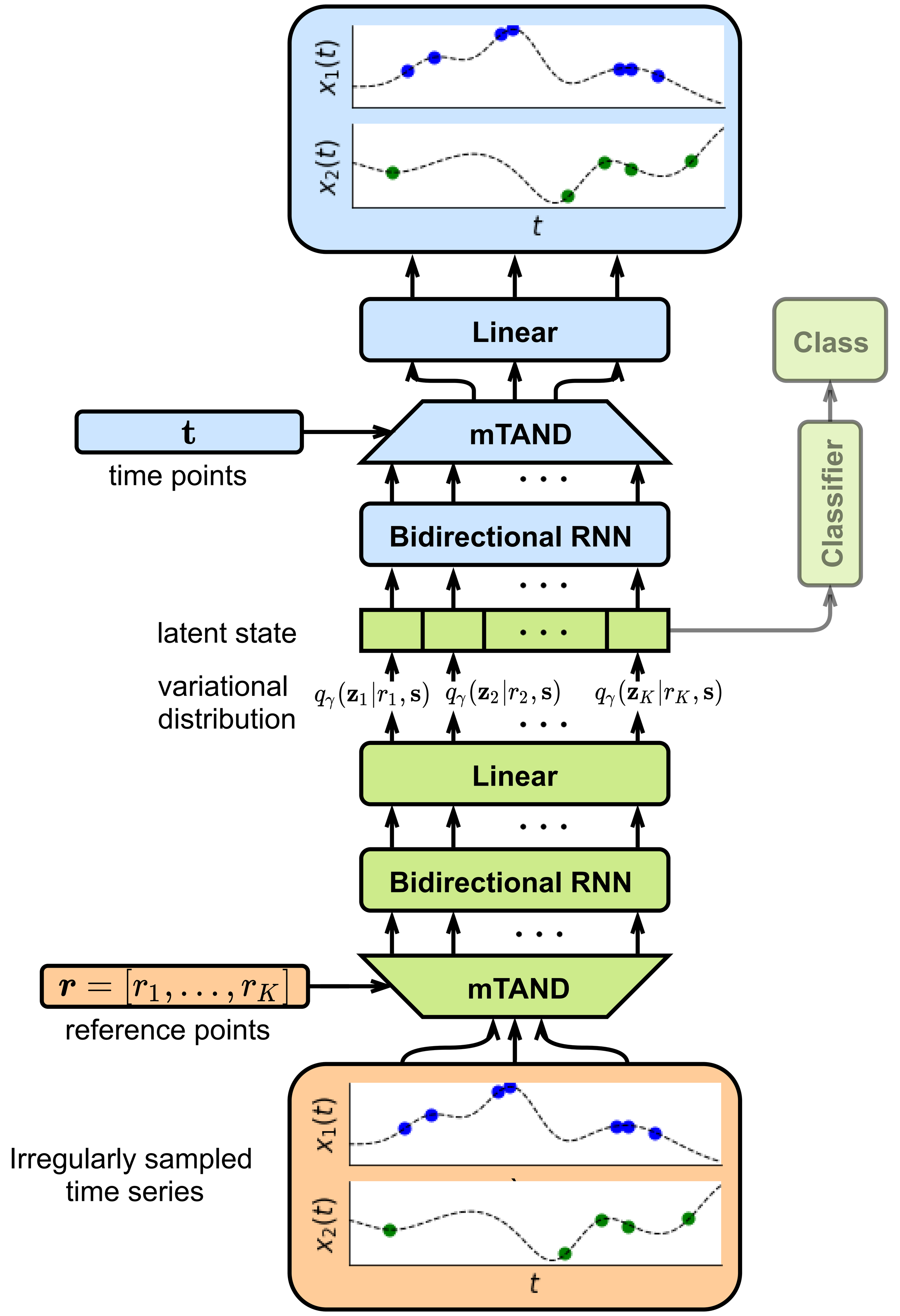


$$\hat{x}_{hd}(t,s) = \sum_{i=1}^{L_d} \mathrm{softmax}\left(\frac{\phi_h(t) w v^T \phi_h(t_{id})^T}{\sqrt{d_k}}\right) x_{id}$$

$$\phi_h(t)[i] = \begin{cases} \omega_{0h} \cdot t + \alpha_{0h}, & \text{if } i = 0 \\ \sin(\omega_{ih} \cdot t + \alpha_{ih}), & \text{if } 0 < i < d_r \end{cases}$$

- **mTANs** produce a fixed dimensional representation by performing a scaled dot product attention over the observed values using the time embedding of the query and key time points.
- **mTAND** materializes mTAN's output at a set of query time points.
- More representational flexibility than previous interpolation models.

Encoder-Decoder Framework



Learning:

- We follow a slightly modified VAE training and maximize the normalized variational lower bound on the log marginal likelihood.

$$\mathcal{L}_{\mathrm{NVAE}}(\theta,\gamma) = \sum_{n=1}^{N} \frac{1}{\sum_d L_{dn}} \Big(\mathbb{E}_{q_\gamma(z|r,s_n)}[\log p_\theta(x_n|z,t_n)] - D_{\mathrm{KL}}(q_\gamma(z|r,s_n)||p(z))\Big)$$

- We augment the model with a supervised learning component that leverages the latent states as a feature extractor.

$$\mathcal{L}_{\mathrm{supervised}}(\theta,\gamma,\delta) = \mathcal{L}_{\mathrm{NVAE}}(\theta,\gamma) + \lambda \mathbb{E}_{q_\gamma(z|r,s_n)} \log p_\delta(y_n|z)$$

- Final predictions of supervised model are computed by marginalizing over the latent variable: $y^* = \arg\max_{y \in \mathcal{Y}} \mathbb{E}_{q_\gamma(z|r,s)}[\log p_\delta(y|z)]$

Experiments

Datasets: Experiments on 3 real-world data sets.

Task	Dataset	Size	Dimension	+ive labels
Interpolation	PhysioNet	8000	41	–
Classification	PhysioNet	4000	41	13.8%
	MIMIC-III	53211	12	8.1%
	Human activity	6554	12	–

Protocols:

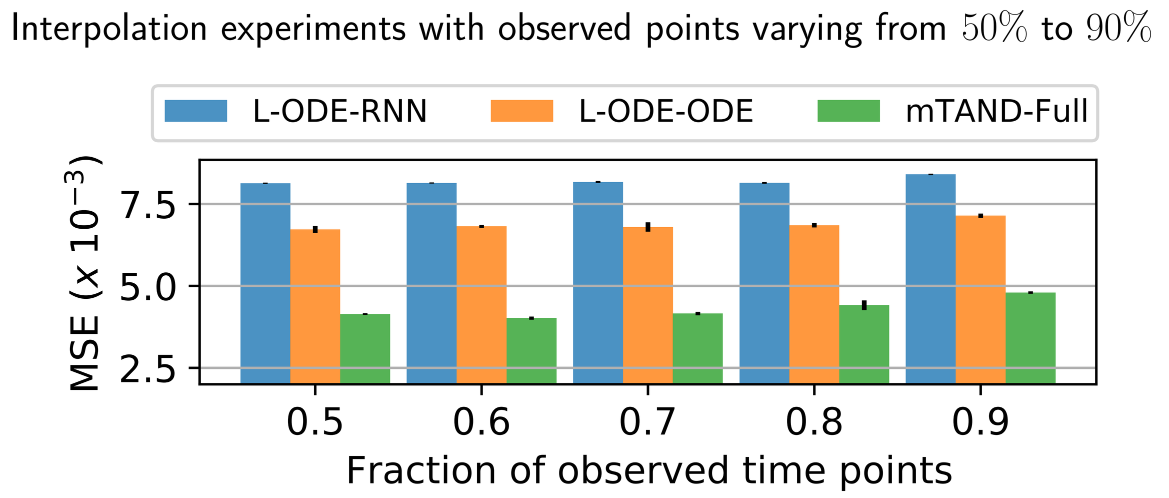
- We randomly divide the data set into a training set (80%) and a test set (20%). We use 20% of the training data for validation.
- In the interpolation task, we condition on a subset of available points and predict values for rest of the time points.
- PhysioNet and MIMIC-III problems are whole time series classification problems.
- Human activity problem is a multi-class classification problem focused on classifying each time point in the time series.
- We repeat each experiment five times using different random seeds to initialize the model parameters.

Models

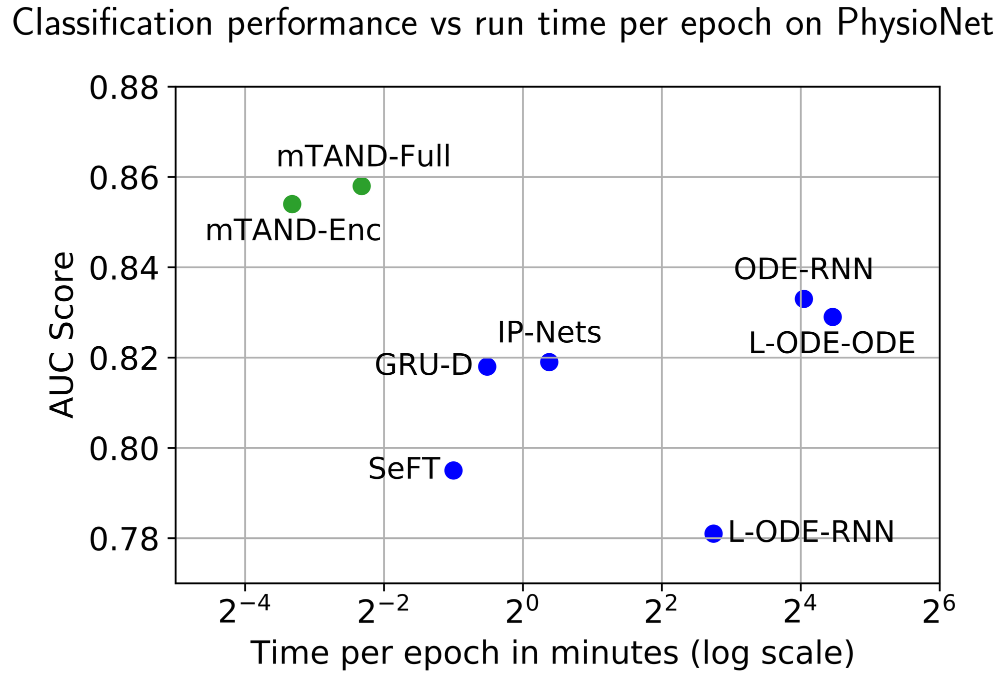
We compare to several baselines and current SOTA methods:

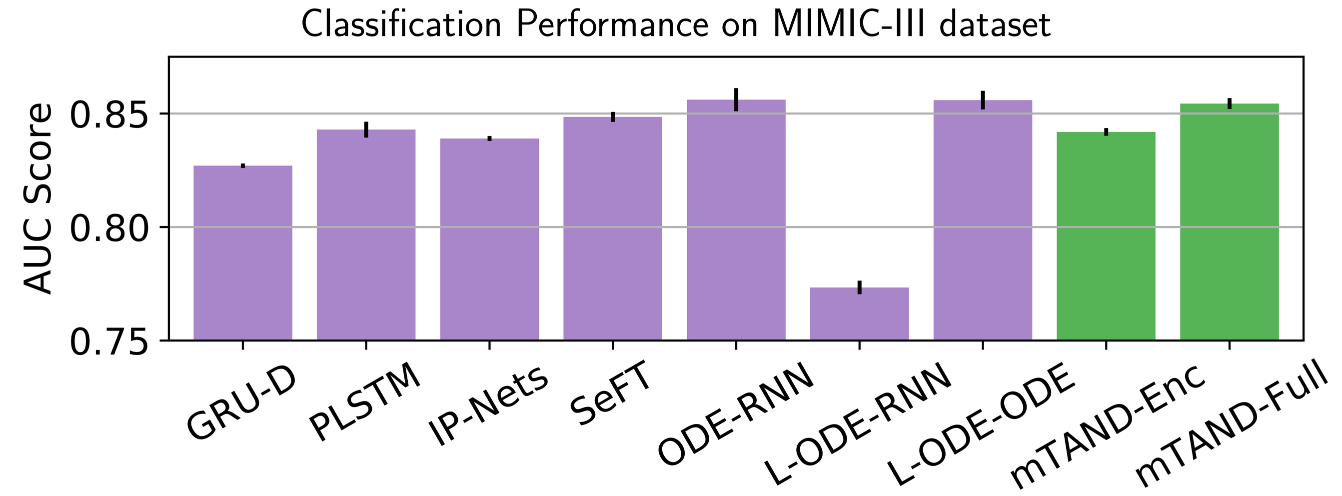
- **GRU-D:** combining hidden state decay with input decay [1].
- **Phased-LSTM:** Captures time irregularity by a time gate [4].
- **IP-Nets:** Interpolation prediction networks followed by a GRU [6].
- **SeFT:** Set function based approach [3].
- **ODE-RNN:** Neural ODE to model hidden state dynamics [5].
- **L-ODE-RNN:** Latent ODE with RNN encoder [2].
- **L-ODE-ODE:** Latent ODE with ODE-RNN encoder [5].

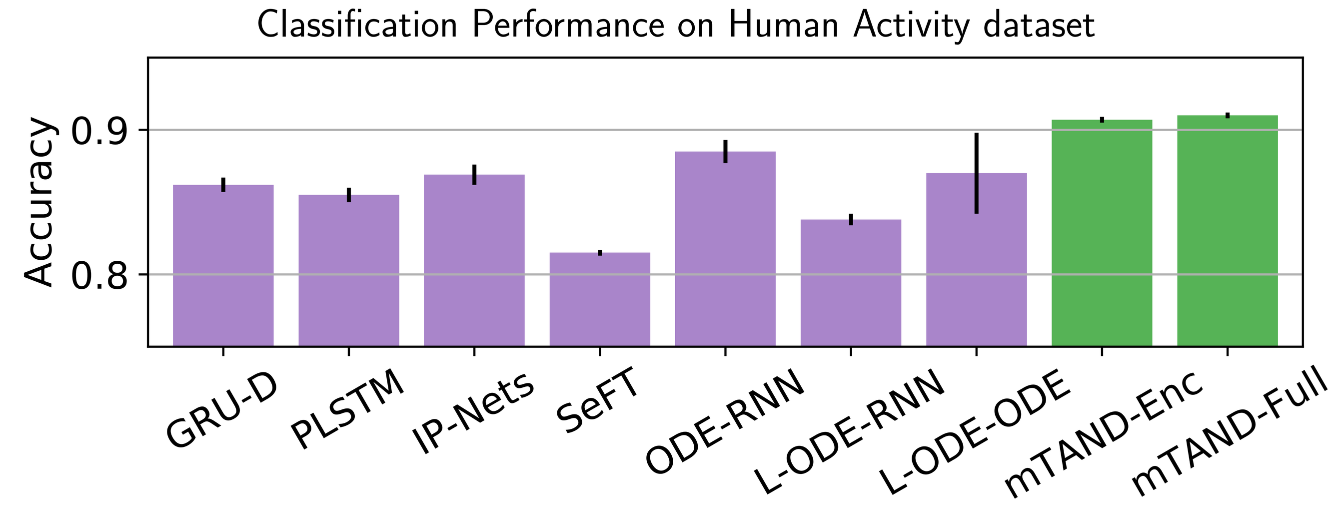
Interpolation Results



Classification Results







References

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

[2] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*. 2018.

[3] M. Horn, M. Moor, C. Bock, B. Rieck, and K. Borgwardt. Set functions for time series. In *Proceedings of the 25th International Conference on Machine Learning*, 2020.

[4] D. Neil, M. Pfeiffer, and S.-C. Liu. Phased lstm: Accelerating recurrent network training for long or event-based sequences. In *Advances in Neural Information Processing Systems*. 2016.

[5] Y. Rubanova, R. T. Q. Chen, and D. K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*. 2019.

[6] S. N. Shukla and B. Marlin. Interpolation-prediction networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2019.