BAYESIAN CONTINUAL IMPUTATION AND PREDICTION FOR IRREGULARLY SAMPLED TIME SERIES DATA

Yang Guo, Jeanette Wen Jun Poh, Cheryl Sze Yin Wong, Savitha Ramasamy Senior Member, IEEE*

Institute for Infocomm Research (I2R), A*STAR, 1 Fusionopolis Way, Singapore 138632

ABSTRACT

Learning from irregularly sampled, streaming, multi-variate time-series data with many missing values is a very challenging task. In this paper, we propose a Bayesian Continual Imputation and Prediction for Time-series Data (B-CIPIT), for learning from a sequence of time-series tasks. First, we develop a Bayesian LSTM based continual learning algorithm, which is capable of learning continually from a sequence of multi-variate time-series tasks, without catastrophically forgetting any representations. Second, we impute missing values in these time-series sequences, in a continual learning setting. We demonstrate and evaluate the robustness of the proposed algorithm on two real-world clinical time-series data sets, namely MIMIC-III [1] and PhysioNet Challenge 2012 [2]. Performance study results show the superiority of the proposed learning algorithm.

***Index Terms*—** Multi-variate Time Series, Continual Learning, Bayesian Long Short-Term Memory, Missing Data Imputation

1. INTRODUCTION

The widespread adoption of IoT and wearable sensors for regular monitoring in various applications such as healthcare and manufacturing has resulted in large volumes of time-series data generation. Analysis of the data has increased the prevalence of AI in these applications. However, owing to the complexity and the non-stationarity of the environment from which the data is generated, offline models that are trained through conventional deep learning techniques are often inadequate to represent the characteristics of the streaming data.

Continual learning methods have recently gained traction due to their ability to adapt the representation of the deep learning models, according to the characteristics of the streaming data [3], without forgetting the past representations catastrophically. To this end, these methods leverage on sharing representations of multiple sequential tasks, while learning and representing unique individual task related distributions efficiently. They achieve this through adapting the architecture of the neural network [4, 5], regularizing the representations of the network [6, 7], or replaying a subset of the past samples or their representations thereof [8, 9]. Additionally, Bayesian inference provides a natural framework for continual learning as it updates the posterior distribution of the trained weights, while learning the likelihood of the oncoming data [10, 11]. Furthermore, the uncertainty estimates of the Bayesian Neural Network (BNN) are used to estimate the significance of weight parameters to preserve important representations of past tasks [12].

However, these methods are largely developed and demonstrated on image data sets, but they are limited in capability for time-series data sets due to the following reasons. Firstly, they are unable to represent the the temporal dependencies of data, which are key characteristics of time-series [13]. Secondly, time-series data are often irregularly sampled due to the varying measurement frequencies of the parameters in the system. For example, in a healthcare setting, there is no control over the frequency of hospital visits by the patients. Moreover, missing data is very common in time-series data sets. While the Bayesian Continual Learning method in [14] has been demonstrated on time-series data sets, they do not continually impute missing values, nor do they address the needs for irregularly sampled time-series data.

Therefore, in this paper, we propose a Bayesian LSTM based continual learning algorithm that is capable of doing both imputation and prediction, as it learns continually from a sequence of multi-variate time-series tasks. It must be noted that the input parameters that are missing due to irregular sampling frequencies are also solved through imputing these values. The proposed algorithm, namely, Bayesian Continual Imputation and Prediction for Irregularly Sampled time-series Data (B-CIPIT) leverages on the inherent ability of variational inference to consider both representations from previous task (previous posterior distribution) and the learning on the current task (the likelihood) without catastrophic forgetting. The CIPIT is evaluated on two real-world multi-variate, time-series data sets, namely, MIMIC-III [1] and Physionet [2], and the evaluation results show that the proposed algorithm helps to learn representations for imputation and pre-

The work is supported by the Industry Alignment Fund Prepositioning (Health Biomedical Sciences) H19/01/a0/023 and the Nanyang Technological University Undergraduate Research Experience on Campus (URECA) Programme.

Guo Yang, Cheryl Wong Sze Yin and Savitha Ramasamy are with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore (e-mail: ramasamysa@i2r.a-star.edu.sg).

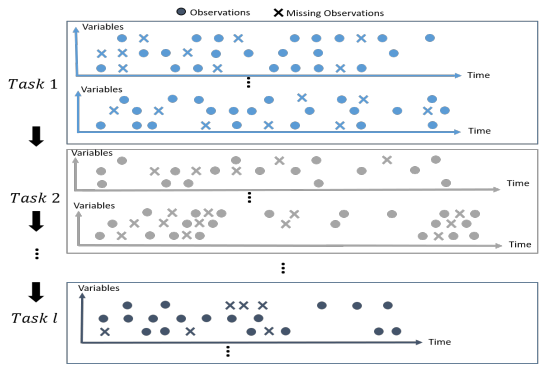

Fig. 1. Illustration of Sequence of Irregularly Sampled Multivariate Time-series Data with Missing values

diction from a sequence of irregularly sampled multi-variate time-series tasks, without catastrophically forgetting any representations.

2. BAYESIAN CONTINUAL IMPUTATION AND PREDICTION FOR TIME-SERIES

In this section, we present the Bayesian Continual Imputation and Prediction for Irregularly sampled time-series data.

2.1. Problem Formulation

Let us assume a sequence of irregularly sampled multi-variate time series data with missing values, as illustrated in Figure 1. Let $(T_1, \ldots T_l)$ be a sequence of l tasks, and the data for each task be $\mathcal{D}_i$; $i = 1, \ldots, l$. Thus, the data for all tasks $\mathcal{D} = [\mathcal{D}_1 \ldots \mathcal{D}_i \ldots \mathcal{D}_l]$ occur sequentially, and let the time-series data for task i be represented by $D_i(\mathbf{X_i}, \mathbf{Y_i})$, where $\mathbf{X_i} \in \Re^{n \times m \times T}$ with n samples, each with m features and T time steps, and $\mathbf{Y_i} \in \Re^{n \times n_c}$ with n samples and n_c classes. The time-series data is characterized by irregular sampling with some of the variables missing at certain time instances. We propose a continual learning algorithm that is capable to address the multiple challenges listed above. We first present the learning algorithm of the Bayesian LSTM, namely, the Bayes by Backprop, and then present the B-CIPIT.

2.2. Bayes by Backprop

The B-CIPIT is based on a Bayesian LSTM (BLSTM), whose weights are learnt as probability distributions through Bayes by Backprop. Let the weights of the B-CIPIT be $\mathbf{w}$ and the data for task i be $\mathcal{D}_i$. The objective is to learn the true posterior distribution of weights $P(\mathbf{w}|\mathcal{D}_i)$. As $P(\mathbf{w}|\mathcal{D}_i)$ is intractable, variational inference is used to approximate the distribution with $q(\mathbf{w}|\theta)$ parameterized by θ and minimizing the Kullback-Leibler (KL) divergence between q and P.

$$\theta^* = \arg\min_{\theta} \mathrm{KL}[q(\mathbf{w}|\theta)||P(\mathbf{w}|\mathcal{D}_i)] \quad (1)$$

Thus, the loss function is derived using Eq. (1):

$$\mathcal{L}(\mathcal{D}_i, \theta) = \mathrm{KL}[q(\mathbf{w}|\theta)||P(\mathbf{w})] - E_{q(\mathbf{w}|\theta)}[\log P(\mathcal{D}_i|\mathbf{w})] \quad (2)$$

Eq. (2) is approximated through M Monte Carlo sampling as:

$$\mathcal{L}(\mathcal{D}_i, \theta) \approx \sum_{j=1}^{M} \log q(\mathbf{w}^{(j)}|\theta) - \log P(\mathbf{w}^{(j)}) - \log P(\mathcal{D}_i|\mathbf{w}^{(j)})$$

2.3. Missing Data Imputation and Outcome Prediction in Continual Learning

Let the input of task i be $\mathbf{X}_i^t$ at time stamp t, $\mathbf{W}_r$ be the weights for the BLSTM, the hidden state $\mathbf{h}^t$ at timestamp t is given by:

$$\mathbf{h}^t = f(p(\mathbf{W}_r|D), \mathbf{h}^{t-1}, \mathbf{X}_i^t) \quad (3)$$

where $p(\mathbf{W}_r|D)$ is the probabilistic distribution of the weights in a BLSTM, which is used to model the uncertainty in f [15]. A temporal decay factor is introduced in the hidden state dynamics [16] to address irregular sampling of inputs, and the missing values in $\mathbf{X}_i^t$ are estimated using:

$$\hat{\mathbf{X}}_i^t = g(\mathbf{h}^{t-1}, \mathbf{X}_i^t) \quad (4)$$

where the imputation function g learns relevant weights to combine history based estimation from h^{t-1} (based on same feature at previous timesteps) and feature based estimation X_i^t (based on other features at same timestep) [17]. A masking matrix $M \in \Re^{n \times m \times T}$ (Eq. (5))

$$M_{i,j,t} = \begin{cases} 0 & \text{if } X_i^t \text{ has a missing value at feature } j \\ 1 & \text{otherwise.} \end{cases}$$

indicates the presence of missing values. Thus the input $\mathbf{X}_i^t$ is updated with $\tilde{\mathbf{X}}_\mathbf{i}^\mathbf{t}$ at timestamp t, according to:

$$\tilde{\mathbf{X}}_\mathbf{i}^\mathbf{t} = (\mathbf{I} - \mathbf{M}) \odot \hat{\mathbf{X}}_\mathbf{i}^\mathbf{t} + \mathbf{M} \odot \mathbf{X}_i^t \quad (5)$$

The missing values of each task are then imputed through optimizing the mean absolute error (MAE) between the actual and predicted values at all time steps $1, \ldots, T$ (L_{imp}), following the imputation loss from [17].

With $\tilde{\mathbf{X}}_i = [\tilde{\mathbf{X}}_\mathbf{i}^\mathbf{1}, \tilde{\mathbf{X}}_\mathbf{i}^\mathbf{2}, \ldots, \tilde{\mathbf{X}}_\mathbf{i}^\mathbf{T}]$ and the LSTM hidden states $\mathbf{H} = [\mathbf{h}^1, \mathbf{h}^2, \ldots, \mathbf{h}^T]$, the predicted output $\hat{\mathbf{Y}}_\mathbf{i}$ is:

$$\hat{\mathbf{Y}}_\mathbf{i} = f_{\text{out}}(\tilde{\mathbf{X}}_i, \mathbf{H}) \quad (6)$$

$$L_{\text{pred}}(\mathbf{Y_i}, \hat{\mathbf{Y}}_\mathbf{i}) = CE(\mathbf{Y_i}, \hat{\mathbf{Y}}_\mathbf{i}) \quad (7)$$

where the classification tasks are learnt through optimizing the cross entropy function (CE) as shown in Eq. (7), and the predicted output $\hat{\mathbf{Y}}_\mathbf{i}$ is obtained using a linear output function (f_{out}).

Thus, given the input $\mathbf{X}_i$ of for a task i (D_i), the imputations and predictions in a time-series data are learnt through

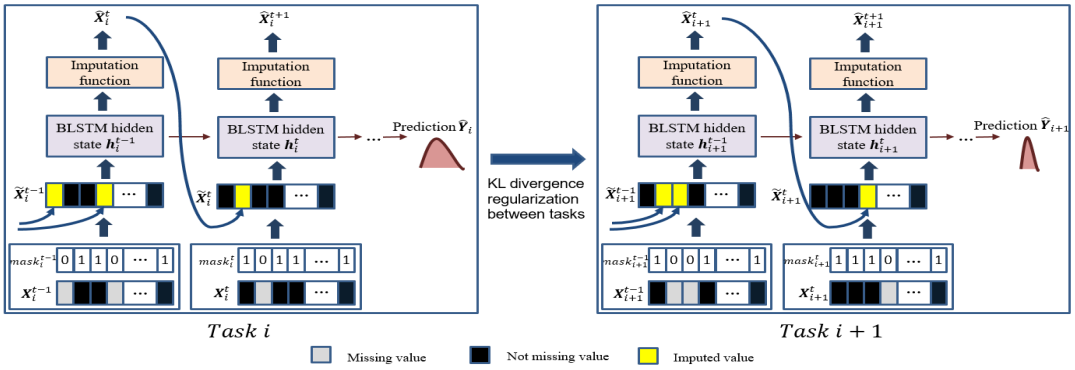

Fig. 2. The Proposed Bayesian Continual Imputation and Prediction for a Sequence of Irregularly Sampled Multi-variate Time-series

estimating the probabilistic distribution of the weights $\mathbf{W}_r$ by optimizing the imputation (L_{imp}) and prediction loss (L_{pred}), recurrently. As the true posterior distribution $p(\mathbf{W}_r|D_i)$ is intractable in general, we approximate it with $q(\mathbf{W}_r|\theta_i)$ using Bayes by Backprop (Section 2.2). Therefore, the total loss function comprises of the imputation errors, the prediction errors and the KL-divergence loss from Bayes by Backprop, as below:

$$\begin{aligned} L_{\text{total}} &= L_{\text{KL}}(q(\mathbf{W}_r|\theta_i), p(\mathbf{W}_r)) \\ &+ L_{\text{imp}}(\mathbf{X}_i, \hat{\mathbf{X}}_i) + L_{\text{pred}}(\mathbf{Y_i}, \hat{\mathbf{Y}}_\mathbf{i}) \end{aligned} \tag{8}$$

where $p(\mathbf{W}_r)$ is the prior distribution for the LSTM weights.

The proposed B-CIPIT has to learn a sequence of tasks continually, which requires preserving representations from previous tasks while adapting network representations to learn new tasks. This is achieved by exploiting the inherent capability of BNN, wherein the prior distribution $p(\mathbf{W}_r)$ of task i is set as the approximated posterior distribution from task $(i-1)$, denoted as $q(\mathbf{W}_r|\theta_{i-1})$. The prior distribution of the first task is set as a zero-centered Gaussian [18]. Thus, continual learning of a sequence of tasks defined by irregularly sampled time series data to impute missing data and predict outcome is accomplished by optimizing the cost function in Eq. (9).

$$\begin{aligned} L_{\text{cont}} &= L_{\text{KL}}(q(\mathbf{W}_r|\theta_i), q(\mathbf{W}_r|\theta_{i-1})) \\ &+ L_{\text{imp}}(\mathbf{X}_i, \hat{\mathbf{X}}_i) + L_{\text{pred}}(\mathbf{Y_i}, \hat{\mathbf{Y}}_\mathbf{i}) \end{aligned} \tag{9}$$

where, the KL divergence between the approximated posterior distributions from the previous and the current tasks $L_{\text{KL}}(q(\mathbf{W}_r|\theta_i), q(\mathbf{W}_r|\theta_{i-1}))$ regularizes the model such that the weight parameters do not deviate too far away from the weight parameters obtained to learn the previous tasks.

In summary, the proposed B-CIPIT optimizes L_{cont} (Eq. (9)) towards (a) Imputation of missing values (including those missing due to irregular sampling) (b) Prediction of outcomes of interest and (c) Continual learning of a sequence of irregularly sampled time series tasks.

3. EXPERIMENTS AND RESULTS

In this section, we evaluate the performance of our B-CIPIT for simultaneous missing data imputation and in-hospital mortality prediction for a sequence of irregularly sampled multi-variate time series tasks, using two publicly available datasets recorded from patients in the intensive care unit (ICU), namely PhysioNet Challenge 2012 (PhysioNet) and the MIMIC -III. For PhysioNet, the input comprises 35 numerical features in time-series samples from 4000 admissions, with 78% missing values. For MIMIC-III, the input comprises 12 numerical features in time-series samples from 14,681 admissions, with 48% missing values. Each input sample in both the data set comprises 48 hourly time steps.

In this paper, the performance of the B-CIPIT is evaluated for its ability to learn continually from a sequence of irregularly sampled time series tasks, where a task is defined as a batch of new patient registrations into the ICU. We define the continual learning problem in the Physionet data set through a sequence of 2 tasks, with each task consisting of 2000 randomly selected admissions. Similarly, we divide the MIMIC -III data into a sequence of 3 tasks, each consisting of equal number (≈ 4985) of randomly selected admissions.

3.1. EVALUATION: Baselines and Metrics

The proposed B-CIPIT is evaluated for its ability to impute and predict continually, for a sequence of time-series tasks, in comparison with the following baselines:

Bayesian Disjoint Imputation and Prediction for Irregularly sampled Time-series (B-DIPIT): Each task is trained on an Independent Bayesian LSTM model. The average performance of the models is reported.

Joint Imputation and Prediction for Irregularly sampled Time-series (JIPIT): A Bayesian LSTM model is trained using data from all tasks, under the assumption that the all task data is available *a priori*. We first compare the performances of different LSTM architectures in the JIPIT framework, viz.,

Table 1. Performance Results of B-CIPIT on Physionet and MIMIC-III Data sets

Performance Results of B-CIPIT on PhysioNet Dataset									
Evaluation Component	Measures	Model						Imputation	
		B-DIPIT	JIPIT			B-FIPIT	B-CIPIT	Forward	Backward
			B-JIPIT	GRU-D	RITS				
Prediction	AUROC	0.83±0.01	0.85 ±0.002	0.828	0.840	0.84±0.001	0.85±0.01	0.80±0.017	0.8±0.03
	AUPRC	0.53 ±0.02	0.56 ± 0.01	-	-	0.55±0.02	0.56±0.02	0.48±0.03	0.48±0.05
Imputation	MAE	0.3±0.0004	0.3±0.0004	0.559	0.300	0.31±0.0008	0.3±0.001	0.93±0.00	0.93±0.00
	MRE	0.43±0.0005	0.43±0.0005	0.776	0.419	0.44±0.001	0.42±0.002	1.32±0.00	1.31±0.00
Catastrophic Forgetting (BWT)	AUROC	-	-	-	-	6e-05±0.015	0.02±0.013	-	-
	AUPRC	-	-	-	-	0.006±0.05	0.15±0.015	-	-
	MAE	-	-	-	-	0.01±0.0001	- 0.0003±0.002	-	-
	MRE	-	-	-	-	0.02±0.002	-0.0004±0.003	-	-

Performance Results of B-CIPIT on MIMIC-III Dataset									
Evaluation Component	Measures	Model						Imputation	
		B-DIPIT	JIPIT			B-FIPIT	B-CIPIT	Forward	Backward
			B-JIPIT	GRU-D	RITS				
Prediction	AUROC	0.78±0.006	0.81±0.006	0.790	0.805	0.78±0.01	0.80±0.01	0.79 ±0.005	0.79±0.0105
	AUPRC	0.42 ± 0.01	0.47±0.01	0.421	0.432	0.41±0.01	0.44±0.02	0.43±0.01	0.42±0.02
Imputation	MAE	0.16±0.0007	0.15±0.0008	0.390	0.151	0.17±0.0013	0.15±0.0009	0.74±0.00	0.74±0.00
	MRE	0.33±0.0014	0.31±0.0016	0.779	0.300	0.34±0.0028	0.31±0.0018	1.55±0.00	1.55±0.00
Catastrophic Forgetting (BWT)	AUROC	-	-	-	-	-0.007±0.003	0.015±0.007	-	-
	AUPRC	-	-	-	-	0.13±0.1	0.012±0.033	-	-
	MAE	-	-	-	-	0.008±0.002	-0.005±0.001	-	-
	MRE	-	-	-	-	0.02±0.004	-0.011±0.001	-	-

with the Bayesian LSTM (B-JIPIT), Gated Recurrent Network with Decay (GRU-D) architecture [19] and the Recurrent Imputation for time-series (RITS) architecture [20]. It must be noted that the JIPIT is the upper bound performance. **Bayesian Fine-tune Imputation and Prediction for Irregularly sampled Time-series (B-FIPIT)**: A Bayesian LSTM model is trained incrementally on sequential tasks, without any strategies for continual learning.

In addition, the imputation performance is evaluated against the **forward** and **backward** imputation methods, within the upper bound of the B-JIPIT framework. The average of AUROC and AUPRC of all tasks are used for prediction evaluation, while MAE and MRE are for imputation. The Backward Transfer (BWT) of B-CIPIT are also reported to evaluate the the level of forgetting during continual learning.

3.2. Performance Results

Table 1 presents the performance of the proposed B-CIPIT, against the baselines, on the Physionet and MIMIC-III data. From the table, it can be observed that the B-JIPIT outperforms the GRU-D and RITS algorithms in the joint learning setting. Hence, our choice of using BLSTM model for continual learning is justified. While using the Bayesian LSTM for continual learning of a sequence of two time series tasks in Physionet, it can be observed that imputation and prediction performances of the proposed B-CIPIT is better than learning the tasks independently through the B-DIPIT, by at least 2%. Moreover, the regularization of weights helps to improve imputation and prediction accuracies by 1%, in comparison to mere fine-tuning in the B-FIPIT. It can also be noted from the BWT that the B-CIPIT does not forget any task catastrophically. In the sequence of 3 tasks in the MIMIC-III data set, the AUPRC of B-CIPIT is better than the B-FIPIT by at least 3%. Additionally, it is shown that the imputation MAE and MRE are much smaller than the traditional forward/backward imputation methods.

4. CONCLUSION

In this paper, we introduce a Bayesian LSTM based continual imputation and prediction for a sequence of tasks defined by irregularly sampled time-series data sets (B-CIPIT). B-CIPIT regularizes its representations across tasks to avoid catastrophic forgetting, while simultaneously optimizing imputation and prediction performances in individual tasks. We demonstrate the continual learning ability of the proposed B-CIPIT by evaluating the Bayesian LSTM model in the absence of continual learning strategy, using two publicly available clinical time series data sets. Performance results show that the proposed B-CIPIT offers better imputation and prediction performance in learning continually from a sequence of time-series tasks.

5. REFERENCES

[1] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[2] S. Farquhar and Y. Gal, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation [Online]*, vol. 101, no. 23, pp. e215–e220, 2019.

[3] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[4] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell, "Progressive neural networks," *CoRR*, vol. abs/1606.04671, 2016.

[5] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang, "Lifelong learning with dynamically expandable networks," in *International Conference on Learning Representations*, 2018.

[6] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell, "Overcoming catastrophic forgetting in neural networks," *CoRR*, vol. abs/1612.00796, 2016.

[7] Friedemann Zenke, Ben Poole, and Surya Ganguli, "Continual learning through synaptic intelligence," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3987–3995.

[8] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim, "Continual learning with deep generative replay," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, pp. 2990–2999, Curran Associates, Inc.

[9] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne, "Experience replay for continual learning," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[10] Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner, "Variational continual learning," in *International Conference on Learning Representations*, 2018.

[11] Noel Loo, Siddharth Swaroop, and Richard E Turner, "Generalized variational continual learning," in *International Conference on Learning Representations*, 2021.

[12] S. Ebrahimi, M. Elhoseiny, T. Darrell, and M. Rohrbach, "Uncertainty-guided continual learning with bayesian neural networks," in *International Conference on Learning Representations*, 2020.

[13] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi, "Unsupervised scalable representation learning for multivariate time series," in *Neural Information Processing Systems*, 2019.

[14] Honglin Li, Payam Barnaghi, Shirin Enshaeifar, and Frieder Ganz, "Continual learning using bayesian neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[15] S. Farquhar and Y. Gal, "A unifying bayesian view of continual learning," in *https://arxiv.org/abs/1902.06494*, 2019.

[16] Z. Ying X. Jun Y. Luo, X. Cai and X. Yuan, "Multivariate time series imputation with generative adversarial networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 1603–1614.

[17] C. Wong Y. Guo and S. Ramasamy, "A bayesian approach for continual learning in clinical time series," in *Machine Learning for Health, 2020*, 2020.

[18] C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner, "Variational continual learning," in *International Conference on Learning Representations*, 2018.

[19] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.

[20] C. Wei, W. Dong, L. Jian, Z. Hao, Y. Li, and L. Lei, "Brits: Bidirectional recurrent imputation for time series," in *Neural Information Processing Systems*, 2018.