PULSEIMPUTE: A NOVEL BENCHMARK TASK AND ARCHITECTURE FOR IMPUTATION OF PHYSIOLOGICAL SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mobile health biophysical sensors that continuously measure our current conditions provide the framework for a personalized guidance system for the maintenance of healthy behaviors. However, this physiological sensor data is commonly plagued with missingness issues that cripple their rich diagnostic utility as well as their ability to enable temporally-precise interventions. While there is a sizable amount of research focusing on imputation methods, surprisingly, no works have addressed the patterns of missingness, quasi-periodic signal structure, and the between-subject heterogeneity that characterizes physiological signals in mHealth applications. We present the PulseImpute Challenge, the first challenge dataset for physiological signal imputation which includes a large set of baselines' performances on realistic missingness models and data. Next, we demonstrate the potential to address this quasi-periodic structure and heterogeneity with our Bottlenecked Dilated Convolution Transformer, a transformer architecture with a self-attention mechanism that scalably increases the temporal receptive field of the query and key transformations. We visually demonstrate that the kernel similarity in the attention model gives high similarity to similar temporal features across quasi-periodic periods, even when the query point has been masked out. We hope the release of our challenge task definitions and baseline implementations will spur the community to address this challenging and important problem.

1 INTRODUCTION

A key approach to improving health outcomes for individuals with chronic conditions like diabetes and heart disease is the use of wearable sensors, e.g. in smart watches, to monitor physiological states and perform diagnosis and intervention. The field of mobile health (mHealth) addresses the challenges of 1) reliably collecting physiological signals, such as ECG and PPG, from wearable sensors during daily life, 2) using machine learning to extract actionable information about health states from sensor data, and 3) delivering mobile interventions to improve outcomes (Rehg et al., 2017). A key challenge is dealing with *missing data*, which arises from multiple causes: insecure attachment of sensors to the body during excessive movement, intermittent wireless dropouts when transmitting sensor data, low battery, and other participant adherence issues (Rahman et al., 2017). Complex patterns of missingness can cause the failure of downstream machine learning methods and result in lost opportunities to identify states of health risk and trigger interventions that could improve outcomes.

While the problem of missing data has been widely studied in statistics (Little & Rubin, 2020) and machine learning (Shukla & Marlin, 2021; Nabi et al., 2020; Mohan et al., 2013), there has been surprisingly little work that systematically addresses the imputation of missing physiological data in the mHealth setting.[1] In contrast, a variety of sophisticated imputation methods have been developed and validated on clinical datasets such as EHR data (Shukla & Marlin, 2021; Rubanova et al., 2019; Yadav et al., 2018). While both mHealth and clinical data contain physiological signals, mHealth

[1]While all deployed mHealth systems contain methods for handling missing sensor data (see (Hovsepian et al., 2015) for a representative example), the approach is typically based on heuristic rules and frequently lacks quantitative evaluation of imputation accuracy.

applications utilize high sampling rates, so that the morphological properties of the signal can be exploited for automated analysis. For example, sampling ECG at 100 Hz exposes the structure of the QRS complex, enabling individual beats to be segmented using peak detection (Park et al., 2017). In contrast, physiological data in an EHR is typically summarized at the level of minutes or hours, which is sufficient for clinical decision-making. As a result, there is a lack of imputation methods designed for *pulsative signals*, which we define as high-frequency physiological signals that exhibit a quasi-periodic structure due to their cardio-pulmonary origin, but also exhibit substantial heterogeneity within- and between-subjects. Since pulsative signals such as ECG (Seshadri et al., 2020), PPG (Biswas et al., 2019), RIP (Geck, 2013), ICG (Patterson, 1989), BCG (Inan et al., 2015), and SCG (Wang et al., 2018) are widely-used in mHealth applications, a systematic approach is needed to develop effective imputation methods.

In this work we introduce *PulseImpute*, a novel benchmark challenge task for the imputation of pulsative signals. The PulseImpute challenge utilizes two public-domain ECG datasets in conjunction with two missingness models that reflect real-world sensor data loss scenarios: *packet loss* due to wireless data transfer and *block loss* that reflects the loss of longer intervals of sampled data due to issues of loose sensor attachment. In addition to quantifying the signal reconstruction accuracy following imputation, we also provide a novel downstream baseline *peak detection* task, that quantifies the effect of imputation on the segmentation of ECG beats. This downstream task captures the impact of imputation on health-related signal analysis. We provide extensive experimental evaluation, benchmarking the accuracy of eight different classical and modern imputation approaches on our novel challenge task.

A key challenge in imputing pulsative signals is to learn representations of the quasi-periodic signal structure that can support the estimation of missing samples. Recently, transformer architectures have been shown to be effective in a self-training task which is based on imputation, making them an attractive choice. While a standard transformer model does not yield good results on pulsative signals, we introduce a new architecture, Bottlenecked Dilated Convolution (BDC) transformer, which is able to effectively capture a larger temporal receptive field while transforming the inputs into the query and key values.

This paper makes the following contributions: 1) Provide a suite of novel benchmarking tasks that focuses on the unique missingness patterns and applications of pulsative mHealth sensor data 2) Introduce our BDC self-attention module for transformers and demonstrate through weight visualizations that it can learn to attend to quasi-periodic features while respecting between-subject heterogeneity 3) Demonstrate the effectiveness of our BDC model against a collection of other classical and deep learning-based imputation models.

2 Related Work

We organize the related works in the following way: 1) Works addressing pulsative data, which are the most closely related, followed by 2) imputation work on non-pulsative but health-related datasets, then 3) other works that have addressed the use of deep learning models for imputation. There are a large number of classical imputation methods and specialized techniques for different data types and problems which we lack the space to review, see Little & Rubin (2020) for an overview. Note that our work assumes an MCAR missingness model, which is an accurate first approximation to the properties of mHealth data, but many prior works have developed more complex missingness models.

Imputation of Pulsative Signals The most closely-related body of work that handles high frequency physiological sensor data is VAR-IM (Bashir & Wei, 2018). They assume stationarity within an ECG time-series and utilize a vector autoregressive model with expectation maximization and prediction error minimization for imputation. Yang et al. (2020) uses RPCA as a matrix completion method for imputation in ECG signals as an intermediate step for arrhythmia classification. Sarker et al. (2016) makes a Missing At Random assumption on their inductive plethysmography and ECG data to utilize a KNN-imputation strategy, however their intervention algorithm is designed to skip over time points that had to be imputed. Similarly, Pires et al. (2020) utilizes a KNN for accelerometer data. These works do not explore deep learning approaches and have not developed a framework for pulsative signal imputation.

Imputation of non-Pulsative Health Data TAR-BRITS (Feng & Narayanan, 2019), is a SOTA deep learning method which uses a bidirectional RNN to impute multivariate health data sampled at a low frequency, such as breathing rate, heart rate, and steps.[2] Wu et al. (2020) utilizes a meta-learning approach with a convolutional autoencoder imputation model on low frequency heart rate data collected from smartwatches, which is extended from previous work in Lin et al.. Cheng et al. uses Gaussian Processes to impute daily total steps and daily sleep hours collected from user-generated mHealth data. Note that a recent study Goldberg et al. (2021) found that 11/36 of surveyed mHealth studies used a multiple imputation method to address missingness.

Widely-used datasets for benchmarking imputation performance include the PhysioNet/CinC Challenge 2012 and MIMIC-III Clinical Database, both of which consist of sparse irregularly-sampled multivariate time-series of vital measurements (Silva et al., 2012; Johnson et al., 2016). The KDD Challenge Cup 2018 is another common dataset, which is regularly sampled and quasi-periodic, but it lacks the distinctive morphology present in longitudinal health sensors data that reflects each individual's cyclic bodily functions (KDD (2018)). In summary, while these prior challenge datasets are health-related, they lack the pulsative properties of the ECG signals in our benchmark.

In summary, prior work on imputation of health-related signals has not addressed the imputation of pulsative signals in a systematic or thorough way, and there is a lack of a well-defined challenge task and benchmarks to spur research progress. Our work addresses this gap.

Deep Learning-Based Imputation Models and Datasets In the area of time-series imputation, the most prior works are SOTA models based on deep learning: DeepMVI (Bansal et al., 2021), BRITS (Cao et al., 2018), and E^{2}GAN (Luo et al., 2019). DeepMVI utilizes a convolutational self-attention mechanism augmented by a parallel kernel regression on a wide array of time-series datasets, such as sales and electricity consumption. BRITS utilizes a bidirectional RNN with a temporal decay factor, and achieving strong results on imputation accuracy tasks for clinical measurements, air quality, and human activity (Cao et al. (2018)). E^{2}GAN utilizes a RNN-based autoencoder generator and discriminator. This architecture is able to achieve SOTA on multiple downstream classification tasks on clinical measurement and weather datasets (Luo et al. (2019)). In contrast to these approaches, we present a novel bottlenecked dilated attention architecture which is a parameter-efficient procedure for generating the large receptive fields that are needed for imputation of pulsative signals. Bansal et al. (2021) is related to this work in that a block-based missingness approach is used to evaluate pattern reconstruction. In contrast, the works Cao et al. (2018); Luo et al. (2019); Choi et al. (2020); Zhang et al. (2021) utilize a missingness model in which every time point independent and randomly-selected observations are masked out.

3 Challenge Description

We propose utilizing electrocardiogram (ECG) signals as the primary task for physiological sensor benchmarking due to their highly representative pulsative nature of other physiological signals. ECG signals typically have up to 12 different leads, in which each lead measures a different direction of this electrical signal propagation through the heart. In the following sections we explore ECG signal's pulsative nature.

3.1 Physiological Sensor Challenges

Here we elaborate on our two challenges associated with physiological sensor data: 1) Non-stationary covariance stemming from its quasi-periodicity and the 2) between subject heterogeneity

Because ECG signals are a measurement of how the heart operates, they are influenced by the biological phenomenon of the variation of heart beats intervals (heart rate variability), that causes its period to constantly change, leading to its "almost periodic" quasiperiodicity (Bohr, 2018).

Between subject heterogeneity can be seen in how the ECG morphology will drastically change depending on the subject's heart health with the relative heights and widths holding clinical significance. Additionally, there is much signal heterogeneity for those even within a given heart condition

[2]Note that such low-frequency mHealth data as heart rate and steps is obtained via the automated analysis of high-frequency raw signals (e.g. PPG and accelerometry). But these works do not address missingness at the level of the high frequency signal.

class. For example, if a patient has left ventricular hypertrophy (LVH), this will increase ventricular depolarization and therefore increase the amplitude of the QRS complex.

3.2 mHealth Missingness Tasks

Rahman et al. (2014) formulates the major data loss factors found within mHealth data: Phone off, sensor off, sensor battery down, attachment loss, loss due to jerks, wireless packet loss, wireless connection loss.

Attachment issues are the most prevalent cause of of missing data within mHealth sensor studies, losing 2.23 hours of ECG data per day during awake hours in one study and 1.58 hours per day in another (Rahman et al., 2014). The most common cause of attachment issues was intermittent loosening, which we seek to model with our **extended signal loss task**. Assuming that after a loosening, the subject will notice and then fix it, this may lead to a missingness of period of 1 to 5 seconds. For our task, we iterate from 1 to 5 seconds, choosing a random part of the signal of that length to mask out. With our time signals lasting 10 seconds, this corresponds to 10%-50% of the signal missing.

Another missingness paradigm is given by wireless packet loss while sensor device is sending information packets to the storage/analysis device. While uncommon, an untimely packet loss may mask out the critical R peak in an ECG signal, thus leading the algorithm to not detect the heart beat. The mobile wearables in Hovsepian et al. (2015) send packets that cover .036 seconds and R waves in leads 1 and 2 last 0.35 seconds at the most, for normal individuals (Pérez-Riera et al., 2016). Therefore for our **packet loss imputation task**, we divide the signal roughly into 0.04 second chunks so that each chunk represents a packet being sent. Then with a given probability from .1 to .5, we will lose a packet, and mask out the corresponding chunk.

Multi-lead ECG wearables are not commonplace in mHealth wearables just yet, but the technology for it is steadily developing. Recently, Hsu et al. (2019) developed a Wearable 12-Lead non-contact electrocardiogram monitoring system. In our PTB-XL dataset, we utilize all 12 leads, and applying missing mechanism to each channel independently to allow models to model both temporal and cross-channel relationships.

3.3 mHealth Downstream Task

R peak detection in ECG is widely used to diagnose heart rhythm irregularities and estimate heart-rate variability (Park et al., 2017). Therefore, it is critically important that during imputation of such ECG signals, the reconstructed signal is able to properly recreate the R in the correct places. As a downstream task, we first run the Christov (2004) peak detection algorithm on the clean waveform, then mask out the waveform according to the missingness strategies above, and rerun the peak detection to flag the peaks that have been lost. Next we apply the various imputation strategies to impute the masked chunks before running the peak detection one last time to calculate the percentage of missing peaks we were able to reconstruct. Peaks were matched with a detection tolerance of .02 seconds.

3.4 Physiological Sensor Datasets

The two datasets that we propose benchmarking on are MIMIC-III Waveforms and PTB-XL. Each of the datasets are cleaned and cut to a 10 second length at 100 Hz. The rest of the paper will now assume the waveforms are 10 seconds with 100 Hz.

MIMIC-III Waveform Database is a companion to the MIMIC-III Clinical Database and is composed 67,830 record sets of various waveforms such as ECG and PPG for approximately 30,000 ICU patients. Due to inter-waveform alignment problems, each lead is treated as a separate waveform, thus lending itself to a univariate imputation problem. After data cleaning, we have 569,598 ECG 1 lead waveforms.

PTB-XL is composed of 21,837 clinical 12-lead ECGs from 18885 patients. This data was annotated by up to two cardiologists for 71 different SCP-ECG statements that cover diagnostic, form, and rhythm statements. It contains a rich coverage of different disease pathologies as well as a large portion of healthy control samples.

3.5 BOTTLENECK DILATED CONVOLUTION TRANSFORMER ARCHITECTURE

In this section we will go over our proposed architecture by first giving background on transformers and previous work on convolutional self-attention. We will then explain our BDC self-attention module in detail before explaining the full architecture and loss function.

3.5.1 BACKGROUND

Tsai et al. (2019) demonstrates how the self-attention module can be reformulated as a kernel regression model in equation 1. From a kernel regression perspective, we can the query points act as evaluation points, the keys act as inducing points, and the values act as a representation of the data. Equation 2 demonstrates how the query and key functions only operate on one temporal point at a time, significantly limiting the information within the kernel comparison.

$$\text{Attention}(x_q|S_{x_k}) = \left(\sum_{x_k} \frac{k(f_q(x_q), f_k(x_k))}{\sum_{x_{k'}} k(f_q(x_q), f_k(x_{k'}))} \right) v(X) \tag{1}$$

with $v(X)$ as the value function and $k(\cdot, \cdot)$ as the similarity kernel between query and key functions, which is an exponential kernel in Vaswani et al. (2017). The query and key functions $f_q(x_q) = x_q W_q$ and $f_k(x_k) = (x_k W_k)^T$ can be defined as 1-D convolutions with kernel size=1.

$$\begin{aligned} f_{q_{t,1:\mathrm{D}}} &= X_{t,1:\mathrm{D}} W^q \\ &= \sum_{s=-\infty}^{\infty} f(t-s) X_{s,1:\mathrm{D}} \text{ where } f(t-s) = \begin{cases} W^q & t-s=0 \\ 0 & \text{elsewhere} \end{cases} \end{aligned} \tag{2}$$

Li et al. (2019) introduced convolutional self-attention, which reformulates the filter function in equation 2 as equation 3, allowing for a larger temporal context to be used within the kernel.

$$f(t-s) = \begin{cases} W^{q_{t-s}} & |t-s| \leq \lfloor \frac{i-1}{2} \rfloor \\ 0 & \text{elsewhere} \end{cases} \text{ where } i > 1 \text{ is the filter size} \tag{3}$$

3.5.2 BOTTLENECKED DILATED CONVOLUTION SELF-ATTENTION

As seen in Fig. 1a)., our increased receptive field attention mechanism is able to infer that the query position corresponds to a PR segment, even when its masked out, then allow for the PR segments in other periods to attend to it. When the query position is shifted to where the R wave would be, the attention focuses on the R waves as well. This attention also enhances imputation performance, as shown by the dotted line.

If we were to use a vanilla transformer with a kernel size of 1, the learned attention weights are not meaningful, which can be seen in Fig. 1c). Previous approaches by Li et al. (2019) and Bansal et al. (2021) utilize a convolution self-attention with a maximum kernel size of 9 and 3 respectively, which is much smaller than an ECG period of 75 time points (given 100 Hz sampling). In Figure 1b) we can visualize the convolutational attention weights learned, which also fails to capture the quasi-periodicity. Additionally, naively increasing the kernel size is not scalable, especially given the increased dimensionality with multiple ECG leads and how other datasets may contain sampling rates up 1000 Hz (Nemcova et al., 2021).

In order to increase our receptive field in a scalable way, we draw from van den Oord et al. (2016) utilize stacked dilated convolutions with bottleneck layers and residual connections. The architecture is shown in Fig. 2. The initial bottleneck layer helps to minimize the effects of high dimensional model allowing for longer and more filters and residual connections facilitate efficient training (He et al., 2015). Our BDC model using this method has 11 million parameters with an effective temporal receptive field of 127, compared to a model that utilizes a convolution with kernel size of 127 having 273 million parameters.

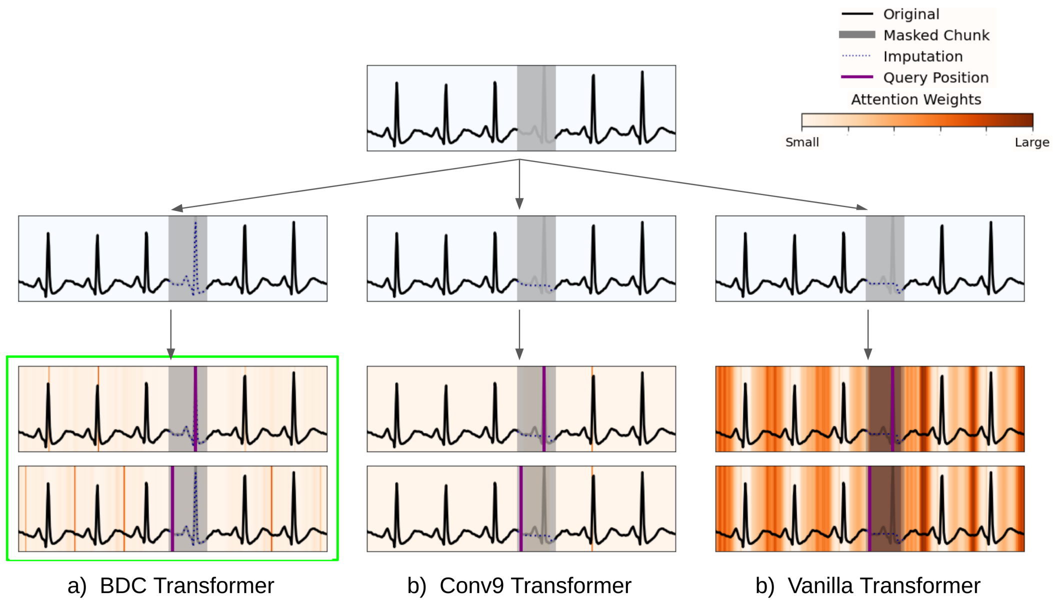

Figure 1: Imputation results with attention weights

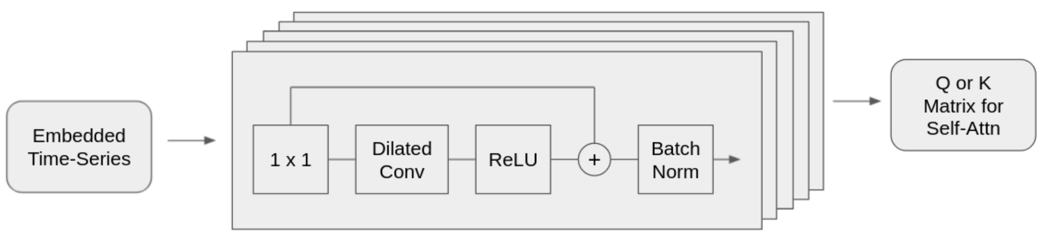

Figure 2: Bottlenecked Dilated Convolution Module. In our model, we have 4 layers with dilated factors of 1,2,2,4

We can now visualize the attention weights in Fig. 1a) learned and see that all of the other key positions corresponding to the query's wave position have an increased attention weights, clearly reflecting the quasi-periodicity of the signal.

3.5.3 Full Architecture

We can will now describe our full architecture. Due to the increased positional context being baked into the self-attention layer, we find that positional embeddings are not necessary for giving positional information. Outside of the transformer module, the rest of the architecture is kept simple to showcase the effectiveness of our modified self-attention layer with a single transformer encoder layer.

For the initial embedding of the raw waveform signal, we utilize a singular 1-D convolution to embed the raw signal to a dimensionality of 1024 with a kernel size of 11, followed by a BatchNorm and ReLU. For the final reconstruction layer, we mirror the embedding with a 1-D convolution and kernel size of 11, projecting the encoded signal into the original dimensionality for evaluation.

3.5.4 Masked Predictive Regression Loss Function

For training we utilize a masked predictive regression loss, which builds upon Devlin et al. (2018) and Jiang et al. (2019). This task is designed to force the model to be robust against scenarios where either a large chunk or a small chunk of data is missing as well as not relying too much on other channels for imputation.

For each waveform there is an equal probability of the 2 scenarios occurring:

1. The time-series is divided into .55 second chunks with a probability of being flagged as .15
2. The time-series is divided into .05 second chunks with a probability of being flagged as .15

Then if a chunk is flagged, with a probability of .8, we will set the value of the window to be 0, with a probability of .1, we will introduce sinusoidal noise to the existing waveform, and with a probability of .1, the window will remain unchanged. Given a multivariate input, then there is an equal probability of flagging all channels together or independently. L1 reconstruction loss from the flagged windows will then be calculated and backpropogated.

4 EXPERIMENTS

4.1 BASELINES

We compare our method against the most relevant classical imputation methods as well modern machine learning methods. Models were trained on 2 Titan V GPUs (for MIMIC) or 2 Titan X GPUs (for PTB-XL) until convergence or after 20 hours, whichever came first. BRIT's training conditions were kept constant from the author's original implementation, and all transformer variants were trained with the Adam optimizer with PyTorch's default parameters.

- **Mean Filling**: Replace missing values with given signal's global mean.This is a naive approach that may lead to biased approaches, even among missing completely at random scenarios (Jamshidian & Bentler (1999)).
- **KNN**: Replace missing value with the k=10 nearest temporal neighbor samples. This is a common strategy used for imputing mHealth data (Sarker et al., 2016).
- **Matrix Factorization**: The time-series is formulated as matrix and factorized into low-rank matrices for imputation. A classical approach that takes advantage of information shared across signal dimensions, but is not be applicable in univariate time-series imputation.
- **BRITS**: Bidrectional RNN architecture with delayed gradients with hidden layer size increased to match total parameters of CLASS transformer (Cao et al., 2018). They achieve state of the art results of multiple multivariate time-series imputation tasks. In order control for parameter size compared to our BDC transformer, the total parameters is 11 million with the hidden layers being set to 1200.
- **Vanilla Transformer**: Vanilla transformer encoder introduced by Vaswani et al. (2017) with convolution attention kernel size of 1. Devlin et al. (2018) showed its potential for usage for word token imputation as a pre-training task. It has 8 million parameters
- **Conv9 Transformer**: Transformer with convolution attentional kernel size of 9, which was first introduced by Li et al. (2019). Bansal et al. (2021) extended this work to use a convolutional attention transformer for imputation on an array of multivariate time-series imputation tasks. The maximum kernel size used in either work is 9. It has 25 million parameters.
- **Conv127 Transformer**: To control for temporal receptive field compared to our BDC transformer, this convolutational attention has a kernel size of 127. It has 273 million parameters.
- **BDC Transformer**: Our proposed transformer architecture which uses the bottlenecked dilated convolution self-attention module to enhance the locality of the attention weights. It has an effective temporal receptive field of 127, with total parameter count of 11 million.

Other distinctive methods that were not benchmarked include E^{2}GAN (Luo et al., 2019), mTAN (Shukla & Marlin, 2021), and RC-VAE+NART (Qi et al., 2020). E^{2}GAN utilizes a GAN architecture with a RNN-based autoencoder generator and RNN-based discriminator. However, GAN-based

Table 1: PTB-XL Multivariate ECG Dataset

Packet Loss (RMSE / % Missing Peaks Reconstructed)								
Prob	Mean	KNN	MF	BRITS	Van.Tran.	C9Tran.	C127Tran.	BDCTran.
0.10	14.36 / 9.76%	6.03 / 62.54%	2.10 / **88.96%**	3.03 / 76.18%	5.21 / 69.43%	3.89 / 73.67%	4.75 / 75.65%	**1.98** / 82.80%
0.20	28.77 / 7.75%	12.87 / 59.23%	5.24 / **86.13%**	7.53 / 70.17%	9.77 / 74.17%	7.39 / 77.02%	9.98 / 76.93%	**3.78** / 85.27%
0.30	43.11 / 6.37%	20.77 / 64.79%	9.92 / 82.17%	13.65 / 13.65%	13.97 / 77.70%	10.63 / 79.39%	16.18 / 77.59%	**5.57 / 86.73%**
0.40	57.51 / 5.27%	30.26 / 64.94%	16.90 / 77.03%	21.04 / 58.44%	18.55 / 79.73%	14.12 / 80.58%	23.63 / 77.46%	**7.46 / 87.85%**
0.50	71.94 / 4.28%	42.60 / 64.76%	27.28 / 69.26%	29.70 / 53.16%	24.11 / 80.44%	18.44 / 80.77%	32.97 / 76.92%	**9.76 / 88.47%**
Extended Signal Loss (RMSE / % Missing Peaks Reconstructed)								
Seconds	Mean	KNN	MF	BRITS	Van.Tran.	C9Tran.	C127Tran.	BDCTran.
1	14.25 / 1.44%	5.65 / 69.94%	**2.14 / 94.00%**	13.37 / 47.81%	4.81 / 79.39%	3.77 / 84.50%	6.25 / 81.07%	2.78 / 92.37%
2	28.72 / 0.65%	12.25 / 70.42%	5.64 / 89.12%	27.35 / 34.30%	8.36 / 85.89%	7.13 / 87.43%	13.09 / 82.96%	**5.28 / 93.23%**
3	43.65 / 0.42%	20.71 / 70.14%	12.19 / 79.22%	38.85 / 26.06%	12.26 / 88.93%	10.96 / 88.77%	21.63 / 83.47%	**7.91 / 93.65%**
4	59.52 / 0.39%	35.62 / 69.45%	24.57 / 60.70%	52.45 / 18.65%	18.95 / 89.14%	17.18 / 87.41%	34.10 / 82.16%	**11.03 / 93.12%**
5	76.15 / 0.67%	59.75 / 59.42%	45.13 / 41.37	69.18 / 13.93%	36.88 / 75.20%	31.58 / 73.17%	51.97 / 67.31%	**21.01 / 81.65%**

Table 2: MIMIC Univariate ECG Dataset

Packet Loss (RMSE / % Missing Peaks Reconstructed)								
Prob	Mean	KNN	MF	BRITS	Van.Tran.	C9Tran.	C127Tran.	BCDTran.
0.10	3.82 / 1.02%	4.22 / 0.95%	**N/A**	3.55 / **0.095%**	3.34 / 0.96%	4.06 / 0.95%	3.31 / 0.92%	**1.04** / 1.10%
0.20	7.66 / 1.18%	8.47 / 1.07%	**N/A**	8.57 / 1.05%	5.80 / 1.10%	8.15 / 1.07%	6.74 / 1.03%	**1.96** / 1.39%
0.30	11.49 / 1.33%	12.70 / 1.17%	**N/A**	12.84 / 1.17%	8.50 / 1.25%	12.22 / 1.19%	10.26 / 1.12%	**2.91 / 1.69%**
0.40	15.32 / 1.45%	16.92 / 1.24%	**N/A**	16.88 / **1.28%**	11.60 / 1.39%	16.29/ 1.30%	13.91 / 1.19%	**4.03** / 1.97%
0.50	19.16 / 1.56%	21.14 / 1.29%	**N/A**	21.09 / 1.41%	15.11 / 1.54%	20.36 / 1.42%	17.75 / 1.27%	**6.35 / 2.29 %**
Extended Signal Loss (RMSE / % Missing Peaks Reconstructed)								
Seconds	Mean	KNN	MF	BRITS	Van.Tran.	C9Tran.	C127Tran.	BCDTran.
1	3.84 / 0.78%	4.24 / 0.77%	**N/A**	4.23 / 0.78%	3.94 / 0.78%	4.07 / 0.78%	4.13 / 0.77%	**3.61 / 0.96%**
2	7.72 / 0.74%	8.45 / 0.73%	**N/A**	8.45 / 0.74%	7.93 / 0.75%	8.14 / 0.74%	8.29 / 0.73%	**6.61 / 0.98%**
3	11.61 / 0.70%	12.66 / 0.69%	**N/A**	12.72 / 0.70%	11.91 / 0.71%	12.19 / 0.71%	12.42 / 0.70%	**10.10 / 0.99%**
4	15.50 / 0.67%	16.87 / 0.66%	**N/A**	17.03 / 0.67%	15.87 / 0.69%	16.30 / 0.68%	16.69 / 0.67%	**14.08 / 0.97%**
5	19.38 / 0.74%	21.10 / 0.72%	**N/A**	21.31 / 0.74%	19.81 / 0.77%	20.38 / 0.75%	20.86 / 0.74%	**17.89 / 0.98%**

imputation networks learn to follow the distribution of the incomplete data rather than the complete data, which would make them not well suited for physiological sensors data. This data has high between-subject heterogeneity, even among those in the same class (i.e. two people with the same heart condition will have different relative waveform heights in an ECG). This requires imputation networks to carefully impute based off of each individual's own signal morphology. mTAN addresses imputation as an interpolation task in an irregularly sampled time-series, which may work well in other sparsely yet regularly sampled time-series, but is not appropriate in our densely regularly sampled time-series. RC-VAE+NART improves on Liu et al. (2019), achieving SOTA results on sequential imputation for trajectories, but this is distinct from our signals domain.

4.2 Results and Discussion

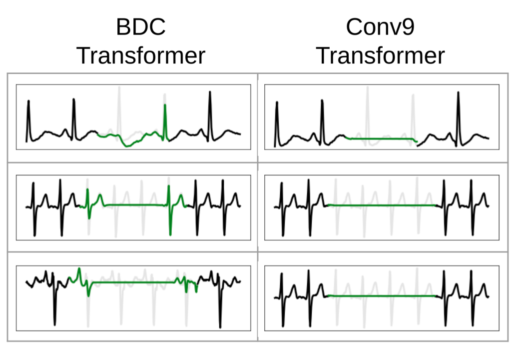

Figure 3: Visualization of Univariate Imputation Results with Increasing Chunk Size

Our BDC Transformer outperforms the other baselines on most of the imputation tasks designed around the mHealth missing data paradigm. Despite this, as we can see in in Fig. 3 and for the percentage of R peaks reconstructed, there is room for improvement in the imputation results in the univariate setting. Matrix factorization does very well in the multivariate imputation problems, able to exploit the cross channel correlations, but it unfortunately cannot be used in the univariate case, as the data matrix starts out being rank 1. Vanilla transformer ends up doing consistently better than the convolutional attention transformer models, likely because of the greatly increased dimensionality from the convolutional attention.

5 Conclusion and Future Work

In summary, we propose a new set of imputation tasks that focus on the unique missingness paradigm present in mHealth pulsative sensors. We show the utility of our BDC Transformer model in utilizing large temporal receptive fields to calculate attention, and its strong performance in these imputation tasks. For future work, we will design additional tasks that utilize further downstream tasks to test how the models impute key signal features, such as a R-R peak detection task and introducing further missingness paradigms such as using accelerometer data to artificially generate missingness.

References

Parikshit Bansal, Prathamesh Deshpande, and Sunita Sarawagi. Missing value imputation on multidimensional time series. *CoRR*, abs/2103.01600, 2021. URL `https://arxiv.org/abs/2103.01600`.

Faraj Bashir and Hua-Liang Wei. Handling missing data in multivariate time series using a vector autoregressive model-imputation (var-im) algorithm. *Neurocomputing*, 276:23–30, 2018.

Dwaipayan Biswas, Neide Simoes-Capela, Chris Van Hoof, and Nick Van Helleputte. Heart Rate Estimation from Wrist-Worn Photoplethysmography: A Review. *IEEE Sensors Journal*, 19(16): 6560–6570, 2019.

Harald Bohr. *Almost periodic functions*. Courier Dover Publications, 2018.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *NeurIPS*, pp. 6776–6786, 2018. URL `http://dblp.uni-trier.de/db/conf/nips/nips2018.html#CaoWLZLL18`.

Li-Fang Cheng, David Stück, Tom Quisel, and Luca Foschini. The impact of missing data in user-generated mhealth time series.

Tae-Min Choi, Ji-Su Kang, and Jong-Hwan Kim. RDIS: random drop imputation with self-training for incomplete time series data. *CoRR*, abs/2010.10075, 2020. URL `https://arxiv.org/abs/2010.10075`.

Ivaylo I Christov. Real time electrocardiogram qrs detection using combined adaptive threshold. *Biomedical engineering online*, 3(1):1–9, 2004.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

Tiantian Feng and Shrikanth Narayanan. Imputing missing data in large-scale multivariate biomedical wearable recordings using bidirectional recurrent neural networks with temporal activation regularization. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2529–2534. IEEE, 2019.

R. Geck. Respiratory Monitoring During Sleep. In *Encyclopedia of Sleep*, pp. 80–83. Elsevier Inc., 2013. URL `http://dx.doi.org/10.1016/B978-0-12-378610-4.00141-8`.

Simon B Goldberg, Daniel M Bolt, and Richard J Davidson. Data missing not at random in mobile health research: Assessment of the problem and a case for sensitivity analyses. *J Med Internet Res*, 23(6):e26749, Jun 2021. ISSN 1438-8871. doi: 10.2196/26749. URL `https://www.jmir.org/2021/6/e26749`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. Cstress: Towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pp. 493–504, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335744. doi: 10.1145/2750858.2807526. URL https://doi.org/10.1145/2750858.2807526.

Chien-Chin Hsu, Bor-Shing Lin, Ke-Yi He, and Bor-Shyh Lin. Design of a wearable 12-lead noncontact electrocardiogram monitoring system. *Sensors*, 19(7):1509, 2019.

Omer T. Inan, Pierre Francois Migeotte, Kwang Suk Park, Mozziyar Etemadi, Kouhyar Tavakolian, Ramon Casanella, John Zanetti, Jens Tank, Irina Funtova, G. Kim Prisk, and Marco Di Rienzo. Ballistocardiography and Seismocardiography: A Review of Recent Advances. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1414–1427, 2015.

Mortaza Jamshidian and Peter M. Bentler. Ml estimation of mean and covariance structures with missing data using complete data routines. *Journal of Educational and Behavioral Statistics*, 24 (1):21–24, 1999. doi: 10.3102/10769986024001021. URL https://doi.org/10.3102/10769986024001021.

Dongwei Jiang, Xiaoning Lei, Wubo Li, Ne Luo, Yuxuan Hu, Wei Zou, and Xiangang Li. Improving transformer-based speech recognition using unsupervised pre-training, 2019.

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*, 3(1):1–9, 2016.

KDD. Kdd cup 2018, Jul 2018. URL https://www.kdd.org/kdd2018/kdd-cup.

Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/6775a0635c302542da2c32aa19d86be0-Paper.pdf.

Suwen Lin, Xian Wu, Gonzalo Martinez, and Nitesh V. Chawla. *Filling Missing Values on Wearable-Sensory Time Series Data*, pp. 46–54. doi: 10.1137/1.9781611976236.6. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611976236.6.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2020.

Yukai Liu, Rose Yu, Stephan Zheng, Eric Zhan, and Yisong Yue. Naomi: Non-autoregressive multiresolution sequence imputation. In *NeurIPS*, 2019.

Yonghong Luo, Ying Zhang, Xiangrui Cai, and Xiaojie Yuan. E2gan: End-to-end generative adversarial network for multivariate time series imputation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, pp. 3094–3100. AAAI Press, 2019. ISBN 9780999241141.

Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In *Proc. Advances in Neural Information Processing Systems (NeurIPS 2013)*, 2013.

Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Full Law Identification In Graphical Models Of Missing Data: Completeness Results. In *Proceedings of the 37th International Conference on Machine Learning (ICML 20)*, volume PMLR 119, pp. 7153–7163, 2020.

Andrea Nemcova, Enikö Vargova, Radovan Smisek, Lucie Marsanova, Lukas Smital, and Martin Vitek. Brno university of technology smartphone ppg database (but ppg): Annotated dataset for ppg quality assessment and heart rate estimation. *BioMed Research International*, 2021, 2021.

Jeong-Seon Park, Sang-Woong Lee, and Unsang Park. R peak detection method using wavelet transform and modified shannon energy envelope. *Journal of healthcare engineering*, 2017, 2017.

R. P. Patterson. Fundamentals of Impedance Cardiography. *IEEE Engineering in Medicine and Biology Magazine*, 8(1):35–38, 1989.

Andrés Ricardo Pérez-Riera, Luiz Carlos de Abreu, Raimundo Barbosa-Barros, Kjell C Nikus, and Adrian Baranchuk. R-peak time: An electrocardiographic parameter with multiple clinical applications. *Annals of Noninvasive Electrocardiology*, 21(1):10–19, 2016.

Ivan Miguel Pires, Faisal Hussain, Nuno M Garcia, and Eftim Zdravevski. Improving human activity monitoring by imputation of missing sensory data: Experimental study. *Future Internet*, 12(9): 155, 2020.

Mengshi Qi, Jie Qin, Yu Wu, and Yi Yang. Imitative non-autoregressive modeling for trajectory forecasting and imputation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Md. Mahbubur Rahman, Rummana Bari, Amin Ahsan Ali, Moushumi Sharmin, Andrew Raij, Karen Hovsepian, Syed Monowar Hossain, Emre Ertin, Ashley Kennedy, David H. Epstein, Kenzie L. Preston, Michelle Jobes, J. Gayle Beck, Satish Kedia, Kenneth D. Ward, Mustafa al'Absi, and Santosh Kumar. Are we there yet? feasibility of continuous stress assessment via wireless physiological sensors. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, pp. 479–488, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450328944. doi: 10.1145/2649387.2649433. URL `https://doi.org/10.1145/2649387.2649433`.

Md. Mahbubur Rahman, Nasir Ali, Rummana Bari, Nazir Saleheen, Mustafa Al'Absi, Emre Ertin, Ashley Kennedy, Kenzie L Preston, and Santosh Kumar. mDebugger: Assessing and Diagnosing the Fidelity and Yield of Mobile Sensor Data. In *Mobile Health: Sensors, Analytic Methods, and Applications*, chapter 7, pp. 121–143. 2017. doi: 10.1007/978-3-319-51394-2.

James M Rehg, Susan A Murphy, and Santosh Kumar. *Mobile Health: Sensors, Analytic Methods, and Applications*. Springer, 2017. doi: 10.1007/978-3-319-51394-2.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ODEs for Irregularly-Sampled Time Series. In *Proceedings Advances in Neural Information Processing Systems (NeurIPS 19)*, pp. 5320–5330, 2019.

Hillol Sarker, Matthew Tyburski, Md Mahbubur Rahman, Karen Hovsepian, Moushumi Sharmin, David H Epstein, Kenzie L Preston, C Debra Furr-Holden, Adam Milam, Inbal Nahum-Shani, et al. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 4489–4501, 2016.

Dhruv R. Seshadri, Barbara Bittel, Dalton Browsky, Penny Houghtaling, Colin K. Drummond, Milind Y. Desai, and A. Marc Gillinov. Accuracy of Apple Watch for Detection of Atrial Fibrillation. *Circulation*, 141(8):702–703, feb 2020.

Satya Narayan Shukla and Benjamin Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=4c0J6lwQ4_`.

Ikaro Silva, George B. Moody, Daniel J. Scott, Leo Anthony Celi, and Roger G. Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. *2012 Computing in Cardiology*, pp. 245–248, 2012.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. *CoRR*, abs/1908.11775, 2019. URL `http://arxiv.org/abs/1908.11775`.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Edward Jay Wang, Junyi Zhu, Mohit Jain, Tien Jui Lee, Elliot Saba, Lama Nachman, and Shwetak N. Patel. Seismo: Blood pressure monitoring using built-in smartphone accelerometer and camera. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 2018-April, 2018.

Xian Wu, Stephen Mattingly, Shayan Mirjafari, Chao Huang, and Nitesh V. Chawla. Personalized imputation on wearable-sensory time series via knowledge transfer. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, CIKM '20, pp. 1625–1634, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3411879. URL https://doi.org/10.1145/3340531.3411879.

Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. Mining electronic health records (EHRs): A survey. *ACM Computing Surveys*, 50(6):1–41, 2018.

Fei Yang, Jiazhi Du, Jiying Lang, Weigang Lu, Lei Liu, Changlong Jin, and Qinma Kang. Missing value estimation methods research for arrhythmia classification using the modified kernel difference-weighted knn algorithms. *BioMed research international*, 2020, 2020.

Ying Zhang, Baohang Zhou, Xiangrui Cai, Wenya Guo, Xiaoke Ding, and Xiaojie Yuan. Missing value imputation in multivariate time series with end-to-end generative adversarial networks. *Information Sciences*, 551:67–82, 2021. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2020.11.035. URL https://www.sciencedirect.com/science/article/pii/S0020025520311233.