



# Forecasting Time Series in Healthcare With Gaussian Processes and Dynamic Time Warping Based Subset Selection

Chetanya Puri , Member, IEEE, Gerben Kooijman, Bart Vanrumste , Senior Member, IEEE, and Stijn Luca 

**Abstract**—Modelling real-world time series can be challenging in the absence of sufficient data. Limited data in healthcare, can arise for several reasons, namely when the number of subjects is insufficient or the observed time series is irregularly sampled at a very low sampling frequency. This is especially true when attempting to develop personalised models, as there are typically few data points available for training from an individual subject. Furthermore, the need for early prediction (as is often the case in healthcare applications) amplifies the problem of limited availability of data. This article proposes a novel personalised technique that can be learned in the absence of sufficient data for early prediction in time series. Our novelty lies in the development of a subset selection approach to select time series that share temporal similarities with the time series of interest, commonly known as the test time series. Then, a Gaussian processes-based model is learned using the existing test data and the chosen subset to produce personalised predictions for the test subject. We will conduct experiments with univariate and multivariate data from real-world healthcare applications to show that our strategy outperforms the state-of-the-art by around 20%.

**Index Terms**—Forecasting, Gaussian processes, machine learning, time series analysis.

## I. INTRODUCTION

**T**IME series forecasting is an extensive field of research for diverse applications with possibilities in economics, physical or environmental sciences, or healthcare. Traditional treatment of time series includes multiplicative methods such as

the auto-regressive integrated moving average model (ARIMA) and its multivariate treatment or state-space models such as the Kalman filter and generalised autoregressive conditional heteroskedasticity (GARCH) process that are additive [1]. These methods are well suited for modelling time series when the data are uniformly sampled. However, as the number of time series in a dataset increases, these methods do not scale well because each time series must be trained individually. Moreover, it is difficult to model the shared temporal patterns across various time series in the whole dataset during training and forecasting.

Modelling real-world healthcare related time series for forecasting is often difficult owing to the limited availability of data due to practical constraints. For example, if a study is conducted only with a small number of participants, then the dataset might not always be a complete representation of a given task. However, limited subjects alone might not be the only issue. For example, if the time series data is sampled at high-frequency, such as accelerometer-based human activity recognition, it is possible to create generalizable, high-performing models even when insufficient subjects are present [2]. If individual time-series are sampled at very low sampling rate from a small number of subjects, the modelling becomes difficult, e.g. modelling daily weight gain over a period of pregnancy. The problem of limited subjects and low sampling frequency is further aggravated when the observed time series, univariate or multivariate, are sporadic in nature, i.e., they are noisy and contain missing values. Few examples include sensor failure, data artifacts in climate time series, or in healthcare use-cases. For example, a patient can skip regular health check-up appointments for intentional or unintentional reasons resulting in multiple missing entries in the electronic health record (EHR) [3]. Furthermore, the individual forecasts must be performed as quickly as possible so that timely interventions can be implemented. This further restricts the availability of the personal data required to learn individual patterns.

Modern deep learning techniques have gained traction in time series forecasting because they can utilise multiple time series from the training data to discover non-linear temporal patterns [4]. However, deep learning models expect huge amounts of training data to learn these patterns [5], [6]. Additionally, state-of-the-art deep learning models for time series forecasting

Manuscript received 2 March 2022; revised 18 August 2022; accepted 9 October 2022. Date of publication 13 October 2022; date of current version 6 December 2022. This work was supported by Marie Skłodowska-Curie through European Union's Horizon 2020 Research and Innovation Programme under Grant 766139. (Bart Vanrumste and Stijn Luca contributed equally to this work.) (Corresponding author: Chetanya Puri.)

Chetanya Puri and Bart Vanrumste are with the e-Media Lab, Campus Groep T, STADIUS, Department of Electrical Engineering, KU Leuven, 3000 Leuven, Belgium (e-mail: chetanya.puri@kuleuven.be; bart.vanrumste@kuleuven.be).

Gerben Kooijman is with the Philips Research, 5656AE Eindhoven, The Netherlands (e-mail: gerben.kooijman@philips.com).

Stijn Luca is with the Department of Data Analysis and Mathematical Modelling, Ghent University, 9000 Ghent, Belgium (e-mail: stijn.luca@ugent.be).

Digital Object Identifier 10.1109/JBHI.2022.3214343

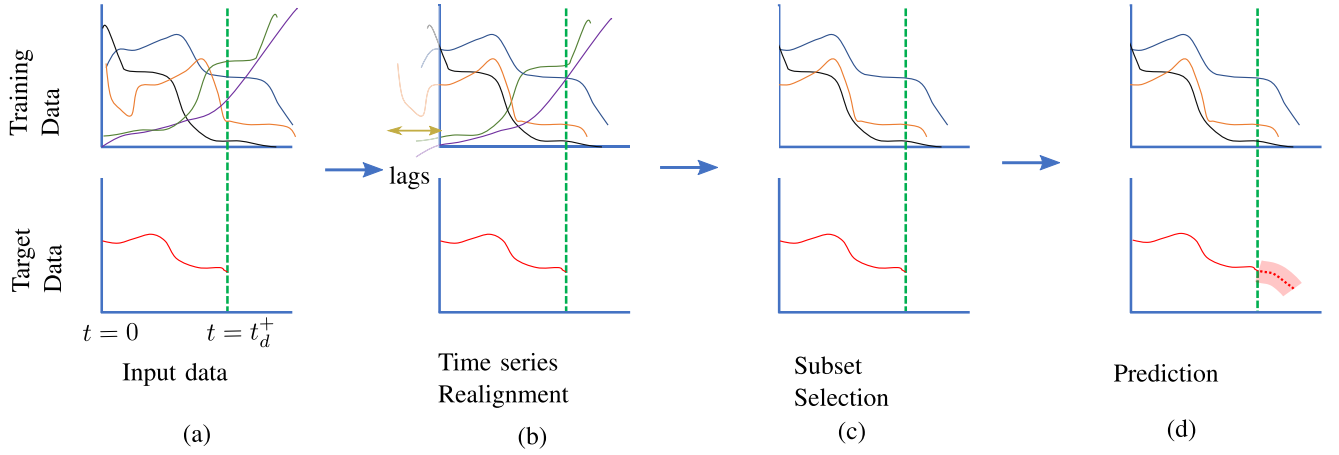


Fig. 1. An example to illustrate our SS-GP approach. (a) The training and target time series that are considered (the green dotted line shows that target data is only available until time ( $t_d^+$ )). (b) The training data that are aligned in time with the target time series. (c) A subset of the training data that share similar temporal characteristics with the target data (the purple and the dark green curves are therefore discarded). (d) The training data and the available target data are used to predict a sequence of future values in the target time series (red dotted line).

still suffer when the training data is sporadic in nature and multi-step forecasting is difficult in the presence of insufficient time series data [7], [8].

This work will develop methods for forecasting time series data in healthcare applications where for each participant time series data is available. Given a target time series (of a test subject) of non-uniformly sampled instances, the main aim is to predict future values over a period of time. In particular, we will treat the following challenges which are often encountered in healthcare data: (a) time series forecasting when for each subject the data are non-uniformly sampled and can have a very low sampling frequency and (b) the estimation of individualised models while very little individual data are available. Note that these data-related difficulties are even more challenging when at the same time the number of subjects is low.

The solution proposed in this paper consists of a subset selection (SS) approach to select time series from the training data (of other subjects) that share temporal similarities with the target time series. This subset of time series is then used to train a non-parametric Gaussian process (GP) in a Bayesian way [9]. Modelling unevenly sampled time series with Gaussian process-based techniques eliminates the need to impute data to make them uniformly sampled. We will show that this approach (further referred to as SS-GP) can improve the target series' forecasting performance, especially when the time series in the selected subset are aligned in time with the target time series.

Fig. 1 showcases an example: first, the training data are aligned with the target time series; second, a subset of time series from the training data is selected that share similar temporal characteristics with the target time series. The subset is then used to train a GP for multi-step ahead prediction, i.e., for predicting a sequence of future values in the target time series.

We experiment with two real-life time-series datasets from healthcare to prove the efficacy of the proposed solution in multi-step time series forecasting. We further demonstrate the implications of limited individual data on training by varying the availability of data in time and assessing the prediction error. We

empirically show that our approach not only reliably predicts in the case of missing observations but also accurately predicts multiple steps ahead in time in the case of limited personal data.

The main contributions of this paper are:

- We propose a new multi-step time series prediction approach that can handle time series with non-uniformly sampled time series data in limited datasets.
- We design a time series realignment technique that tackles time series in a training set that were initiated at different times. In other words, when time  $t_0$  of the training time series is different, realigning them with respect to each other prior to modelling leads to a more exact pattern match and a more precise forecast.
- We suggest dynamic subset selection, which takes advantage of shared temporal patterns to dynamically select a smaller subset of time-series from the training data.
- Finally, we empirically show that the SS-GP approach outperforms state-of-the-art approaches on two real-world healthcare datasets where there is a need to predict early and where missing data are inevitable.

## II. RELATED WORK

Time series literature consists of widespread approaches for forecasting ranging from classical works from the 1960 s to contemporary works [10], [11]. Classical works like state-space or autoregressive approaches such as ARIMA for univariate and VARIMA for multivariate approaches exist that predict the individual observations in time series [12]. Much of these approaches are applied in an auto-regressive manner where one step predictions are achieved by applying the learned model recursively. This tends to achieve significant errors in prediction if the forecast horizon is large. Currently, deep learning-based methods such as recurrent neural networks (RNNs) are popular due to their automatic feature extraction abilities in sequence modelling. Improved variants of RNNs that alleviate vanishing gradient problems such as long-short term memory networks [4]

and gated recurrent units (GRU) [13] are capable of capturing long term dependency with uniformly sampled sequence data. Authors in [14] create a mask where the data is missing and use this mask along with available data as input thus utilising missingness in data as informative features to train RNNs and cope with missingness in the data.

Multiple approaches in deep learning have focused on time series classification and regression in healthcare ranging from ECG classification [15] to glucose forecasting [16]. Authors in [14], [17] have presented works that are able to diagnose a condition, such as sepsis in an intensive care unit environment, by learning from multivariate clinical data using resources such as electronic health records (EHRs). The majority of these methods that can manage missing data have been trained on a significant amount of data, providing them an advantage. However, when insufficient training data is available, traditional machine learning strategies outperform deep learning strategies [18]. There have not been any systematic work that handles limited data availability. We attempt to address such deficiency in training data that stems from either (a) the irregularly sampled time series, or (b) the limited number of samples of an individual time series resulting from the necessity to predict as soon as possible.

Gaussian processes (GPs) provide a framework to model time series in the presence of such irregularly sampled instances and can quantify the uncertainty of predictions. For example, GP models are used in clinical time series classification and imputation [19].

This work proposes a personalised approach for multi-step time series forecasting that can handle non-uniformly sampled time series through Bayesian learning.

### III. NOTATION

Let us assume,  $N$  subjects are studied and the *training data* consists of time series data of  $K$  predictor variables denoted by  $x$  at time  $t$  for each subject  $1 \leq j \leq N$ :

$$x_1^j(t), \dots, x_K^j(t).$$

Our goal is to make predictions about a response variable  $y^j(t)$  based on such feature data. For each subject however the time series are sampled at  $i$  different times,  $t_i^j$ , such that,

$$t_1^j < t_2^j < \dots < t_{m_j}^j,$$

where  $m_j$  denotes the number of measurements of the feature  $x_k^j$  ( $1 \leq k \leq K$ ) that are available for the  $j^{th}$  subject. Remark that, for a given subject  $j$ , all predictor variables are measured at the same time instances.

In what follows, the feature data is denoted in matrix notation:

$$\mathbf{X}^j = [x_k^j(t_i^j)]_{ik}$$

denoting a  $m^j \times K$  matrix of which the  $k^{th}$  column contains the data of the  $k^{th}$  feature of the  $j^{th}$  subject over all the time instances.

The measurements of the responses of a subject are collected in a vector:

$$\mathbf{y}^j = [y^j(t_1^j) \dots y^j(t_{m_j}^j)].$$

Note that the response variable for subject  $j$  is sampled at the same time instances as the predictor variables of subject  $j$ .

There are two ways that data in time series might go missing:

- *missing observations within a time-series*: time series in the  $j^{th}$  instance of training or target data might not be evenly spaced, i.e.,  $t_{(i+1)}^j - t_i^j \neq \lambda, \forall i \in \{1, 2, \dots, m^j - 1\}$ , where  $\lambda > 0$  is some constant.
- *missing observations in different time instants within all time-series*: Time series data of predictor variables are not sampled at the same times across different subjects, i.e.,  $t_i^j$  is not necessarily equal to  $t_i^{j'}, \forall j, j' \in \{1, \dots, N\}, i \in \{1, \dots, m^j\}$

Suppose, we are interested in predictions for a *target subject* (indexed with '+') based on the measurements of the predictor variables  $\mathbf{X}^+$  and the measurements of the response variable available up to some time  $t_d^+$ :

$$\mathbf{y}^+ = [y^+(t_1^+) y^+(t_2^+) \dots y^+(t_d^+)],$$

where we assume that  $t_d^+ \ll t_{m^j}^j, \forall j \in \{1, \dots, N\}$  i.e., the available temporal information for a target subject is limited compared to the number of time instances that are available for training for other subjects primarily due to the need for early prediction. The objective is to try to learn a function  $f$ , such that, the future response value at  $h^{th}$  time-step can be predicted as,

$$y_{(d+h)}^+ = f\left(\underbrace{\mathbf{X}^j, \mathbf{y}^j}_{\text{training data}}, \underbrace{\mathbf{X}^+, \mathbf{y}^+}_{\text{target data}}\right) + \epsilon_h, \quad (1)$$

where

$$\epsilon_h \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

is independent and identically distributed (i.i.d) gaussian.

There are two multi-step forecasting strategies, direct vs iterative. Note that we use a direct multi-step prediction strategy where the responses at  $t_{d+1}^+, \dots, t_{d+h}^+$  time steps are predicted using only the available data until time  $t_d^+$ . However, an iterative multi-step forecasting technique predicts only the next time occurrence at  $t_{d+1}^+$  at a time. Multi-step predictions then can be made by including the previously predicted value ( $y^+(t_{d+1}^+)$ ) of the response variable in the training data to predict the response at the next time instance and so on until  $h^{th}$  time-step is predicted [20].

### IV. STATE-OF-THE-ART

In this section, we provide a brief overview of the existing techniques.

#### A. Subset Selection

Despite its simplicity, the  $k$ -nearest neighbours technique remains the benchmark for the classification of univariate time series [21]. In the case of multivariate time series, we employ  $k$ -means based clustering to create  $k$  profiles among the given dataset of time series grouping them by similar patterns. We further discuss the implementation details in Section VI-B2.

## B. Time Series Forecasting

**Maximum Likelihood Estimation (MLE):** A  $p^{th}$ -order polynomial can be estimated with coefficients  $\beta = [\beta_0 \beta_1 \dots \beta_p]^T$  such that  $y^+(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \dots + \beta_p t^p$ . The training can be done by maximizing the likelihood over the available responses  $\mathbf{y}^+ = [y^+(t_1^+) y^+(t_2^+) \dots y^+(t_d^+)]$ ,  $\ell(\mathbf{w}) = P(\mathbf{y}^+|\beta)$ ,

$$\hat{\beta}_{MLE} = \underset{\beta}{\operatorname{argmax}} P(\mathbf{y}^+|\beta) = \prod_{i=1}^d p(y^+(t_i^+)|t_i^+; \beta). \quad (2)$$

(2) is the model created using only a few observations from the target data up to the time  $t_d^+$  days. This method results in personalised models and predictions, but the limited availability of data can hamper inference. This article will show how to properly use data from other subjects to address this issue.

**Maximum-a-posteriori estimation (MAP) [22]:** The maximum likelihood estimate of  $\hat{\beta}$  may be found using the available training data (of other subjects). As an a-priori estimate, the distribution of these coefficient estimates,  $p(\beta)$ , obtained from the  $N$  participants in the training data may be used. The maximum-a-posteriori estimate of the coefficients,  $p(\beta|\mathbf{y}^+)$  is calculated by combining the likelihood learned from the target data with the *prior* distribution learned from the training data using Bayes theorem:

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmax}} p(\beta|\mathbf{y}^+) = \frac{P(\mathbf{y}^+|\beta)p(\beta)}{P(\mathbf{y}^+)}. \quad (3)$$

At time  $t_m^+$ , the prediction is given by  $\hat{\beta}_{MAP}[t_m^+ t_m^{+2} \dots t_m^{+p}]^T$ . In both MLE and MAP, the parameter  $p$  is selected based on the application of interest, which should be known in advance.

**ARIMA:** is a method for forecasting time series data based on correlations in historical data [12]. Time series samples must be consistently spaced when utilising ARIMA algorithms for forecasting. Personal training data can be made uniform using linear interpolation between samples. For a uniformly sampled target time series response variable, an ARIMA model of order  $(p, d, q)$  capable of modelling  $\mathbf{y}_+ = [y^+(t_1^+) y^+(t_2^+) \dots y^+(t_d^+)]$  is defined by the equation:

$$\phi(B)(1 - B)^d y^+(t) = \theta(B)w(t), \quad (4)$$

where  $y^+(t)$  and  $w(t)$  represent time series and random error at time  $t$  respectively.  $B$  is a backward shift operator defined by  $By^+(t) = y^+(t - 1)$ ,  $d$  is the order of differencing.  $\phi(B)$  and  $\theta(B)$  are autoregressive (AR) and moving averages (MA) operators of orders  $p$  and  $q$ , respectively, and are defined as,

$$\begin{aligned} \phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p \\ \theta(B) &= 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q, \end{aligned} \quad (5)$$

where  $\phi_1, \phi_2, \dots, \phi_p$  are the autoregressive coefficients and  $\theta_1, \theta_2, \dots, \theta_q$  are the moving average coefficients.

**LSTM:** Long Short-Term Memory (LSTM) networks are a particular case of Recurrent Neural Networks (RNN) with the ability to model temporal dependencies from the past and have shown outstanding prediction performance [4]. This is done by using *forget*, *memory* and *output gate* that control the flow of the

data during learning. This makes it easier to decide whether the data in each LSTM cell should be discarded, filtered, or added to the next cell [4].

**Gaussian Processes:** The Gaussian Processes (GP) are non-parametric models appropriate for sparsely available data. GP is a collection of random variables, such that the joint distribution of every finite set of them is Gaussian (multivariate) [9]. We are given a training data  $\mathbf{X}$ s for  $N$  subjects:<sup>1</sup>

$$\mathbf{X}\mathbf{s} = \begin{bmatrix} \mathbf{X}^1 \\ \mathbf{X}^2 \\ \vdots \\ \mathbf{X}^N \end{bmatrix} = \begin{bmatrix} x_1^1(t_1^1) & x_2^1(t_1^1) & \dots & x_K^1(t_1^1) \\ x_1^1(t_2^1) & x_2^1(t_2^1) & \dots & x_K^1(t_2^1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1^1(t_{m^1}^1) & x_2^1(t_{m^1}^1) & \dots & x_K^1(t_{m^1}^1) \\ x_1^2(t_1^2) & x_2^2(t_1^2) & \dots & x_K^2(t_1^2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1^2(t_{m^2}^2) & x_2^2(t_{m^2}^2) & \dots & x_K^2(t_{m^2}^2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1^N(t_{m^N}^N) & x_2^N(t_{m^N}^N) & \dots & x_K^N(t_{m^N}^N) \end{bmatrix}, \quad (6)$$

and  $\mathbf{y}\mathbf{s} = [\mathbf{y}^1 \mathbf{y}^2 \dots \mathbf{y}^N]^T$ .  $f$  is defined from (1) as  $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$  with mean and covariance functions  $m(\mathbf{x})$  and  $k(\mathbf{x}, \mathbf{x}')$  respectively. The covariance function encodes all the assumptions of the data such that two independent observations closer to each other have similar outputs. This nearness is used to model the structure of the multivariate time series, given that the covariance remains positive semi-definite [9]. We chose a squared exponential covariance function based on the assumption that the data have independent and identically distributed gaussian noise with variance  $\sigma_n^2$ ,

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2l^2}|\mathbf{x} - \mathbf{x}'|^2\right) \quad (7)$$

Given  $\mathbf{y}\mathbf{s} = [y^1(t_1^1) \dots y^1(t_{m^1}^1) \dots y^N(t_1^N) \dots y^N(t_{m^N}^N)]^T$  and  $\mathbf{K}$  as a matrix  $K_{ab} = k(\mathbf{x}_a, \mathbf{x}_b)$ ,  $\forall \mathbf{x}_a, \mathbf{x}_b \in \mathbf{X}\mathbf{s}$  using (7), and following the optimisation procedure from [9], the hyperparameters  $\{\sigma_f, l, \sigma_n\}$  are estimated by maximising the marginal likelihood  $p(\mathbf{y}\mathbf{s}|\mathbf{X}\mathbf{s}; \{\sigma_f, l, \sigma_n\})$ . The prediction at time  $t_{m^+}^+$  for the observation  $\mathbf{x}_{m^+} = [x_k^+(t_{m^+}^+)]_{m^+k}$  is given by the mean function,  $\mu$  and variance function,  $\sigma^2$ ,

$$\begin{aligned} \mu_{\mathbf{x}_{m^+}} &= \mathbf{k}_+^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}\mathbf{s} \\ \sigma_{\mathbf{x}_{m^+}} &= k(\mathbf{x}_{m^+}, \mathbf{x}_{m^+}) - \mathbf{k}_+^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_+ \end{aligned} \quad (8)$$

where  $\mathbf{k}(\mathbf{x}_{m^+})$  is denoted as  $\mathbf{k}_+$ , and  $\mathbf{k}(\mathbf{x}_{m^+}) = [k(\mathbf{x}_{m^+} \mathbf{x}_1^1) \dots k(\mathbf{x}_{m^+} \mathbf{x}_{m^N}^N)]^T$ .

**Autoregressive Gaussian Processes (AR-GP) [20]:** Peterson et al. [20] employ auto-regressive Gaussian processes (AR-GP) to predict the cognitive decline of Alzheimer's disease patients over the next four time steps. This is further discussed in Section VI-B2. They start by building a population-level forecast model using data from training subjects. They use domain-adaptive GPs to sequentially adapt the GP posterior for the

<sup>1</sup>The superscript represents the  $j^{th}$  subject and not the exponent.



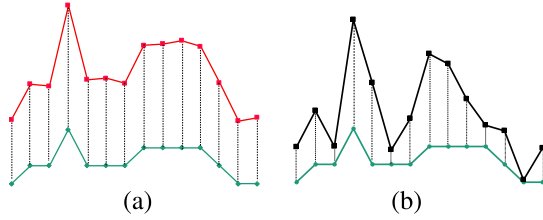


Fig. 2. Normalised Euclidean distances between (a) similar time series and (b) dissimilar time series. The reference time series that is considered is shown in dark green.

test subject using the available data from the test subject. In contrast to our direct technique for multi-step prediction, this is accomplished via an iterative strategy by utilising the data up until time  $t - 1$  to predict the response at time  $t$ . The predictions made are then used again with training data to predict time instant  $t + 1$  and so on.

## V. METHODOLOGY

In this section, the subset selection (SS) based gaussian process (GP) approach (SS-GP) is introduced. First, a novel approach for SS is described. Second, we develop an algorithm to align the time series in the subset with a target time series.

### A. Dynamic Subset Selection

Given a discrete time series  $\mathbf{y}^{ref}$ , and a collection of  $N$  time series  $\mathbf{y}^j$  ( $1 \leq j \leq N$ ) we want to find a time series  $\mathbf{y}^{sim} \in \mathbf{y}^j$  that is closest to  $\mathbf{y}^{ref}$ , i.e.  $dist(\mathbf{y}^{sim}, \mathbf{y}^{ref}) < dist(\mathbf{y}^j, \mathbf{y}^{ref}) \forall j \in \{1, N\}$  [23]. The closeness is calculated by matching time points in two time series based on a distance metric  $dist$ . For example, to calculate the Euclidean distance between two *equal-length* time series  $\mathbf{y}^p = [y_1^p, y_2^p, \dots, y_m^p]$  and  $\mathbf{y}^q = [y_1^q, y_2^q, \dots, y_m^q]$  a *one-to-one* matching is performed to calculate the distance as  $dist(\mathbf{y}^p, \mathbf{y}^q) = \sqrt{\sum_{t=1}^m (y_t^p - y_t^q)^2}$ . Fig. 2 shows examples of *one-to-one* time-point matching with Euclidean distance (dotted line) with Fig. 2(a) exhibiting more similarity with a Euclidean distance of 0.1 as compared to Fig. 2(b) that has a Euclidean distance of 5.1 compared to a reference time series.

1) **Distance Measurement:** Remember that our goal is to make predictions of the response variable  $y^+(t)$  for  $t > t_d^+$ . Our aim in this section is to find a subset of response variables  $y^j(t)$  ( $1 \leq j \leq M$ ) that show similar temporal characteristics with  $y^+(t)$  for  $t < t_d^+$ . This will lead to a subset  $\hat{\mathcal{X}} = \{(\mathbf{X}^1, \mathbf{y}^1), \dots, (\mathbf{X}^M, \mathbf{y}^M)\}$  with  $M \ll N$  of the training dataset  $\{Xs, ys\}$  that is used in a non-parametric GP approach for predicting  $y^+(t)$ . For this purpose, we start by calculating the distances between target response time series data  $\mathbf{y}^+ = [y^+(t_1^+) \ y^+(t_2^+) \ \dots \ y^+(t_d^+)]$  and training data's response variable (nearest to the allowed time point, i.e. ' $\leq t_d^+$ '). Let's denote this distance vector as  $\Omega_+ = [\omega_{1+} \ \omega_{2+} \ \dots \ \omega_{N+}]^T$ , where  $\omega_{j+} = dist([y^j(t_1^+) \ \dots \ y^j(t_d^+)], [y^+(t_1^+) \ \dots \ y^+(t_d^+)])$ . In contrast to equal-length time series in Fig. 2, it is difficult to determine the Euclidean distance (dissimilarity) between two time series with unequal lengths. Therefore, we use Dynamic time warping (DTW) [24] as a distance metric  $dist$  in our study

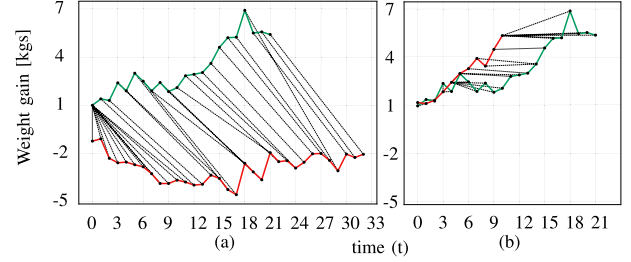


Fig. 3. DTW distances between time series with different lengths. The matched points are indicated by a dotted line. The reference time series is shown in dark green. In (a) the DTW distance is 170 and the time series are more dissimilar than in (b) where the DTW distance is 6.9.

that allows *one-to-many* matching and thus subsumes Euclidean distance. DTW distance has an ability to match time series of different lengths and is robust to shifting and scaling along the time axis [25]. It matches two time series by (i) calculating a local cost matrix between each pair of elements between these time series, and then the goal of minimising the overall cost (distance) is achieved by (ii) finding an optimal alignment that runs along a low cost "valley" within the cost matrix [26]. Fig. 3 illustrates that DTW first aligns the time series. Points of the time series that are matched are connected by a dotted line. The final distance is computed by taking the sum of the Euclidean distances between the matched points. Clearly, the reference time series (in green) is more similar in trend to the time series shown in Fig. 3(b) compared to the one shown in Fig. 3(a).

Since we are calculating the DTW distances in the output space, i.e., between the response time series' ( $\mathbf{y}^j$ ), the distance measurement is applicable in settings where the input time series is multivariate. As long as the output time series is univariate the DTW distance can be calculated as proposed, which is the case in many healthcare applications. For a multidimensional DTW treatment, the reader is referred to [27].

2) **Subset Selection:** After calculating the distance vector  $\Omega_+$  of length  $N$  between a target time series  $\mathbf{y}^+$  and other time series' ( $\mathbf{y}^j$ ), the nearest subjects are determined by dynamically calculating a cut-off point for the target time series in the following way:

- i) *Arrange elements by their closeness to the target time series:* sort the distance vector  $\Omega_+$  in increasing order as  $\hat{\Omega}_+ = [\hat{\omega}_{1+} \ \hat{\omega}_{2+} \ \dots \ \hat{\omega}_{N+}]^T$ , such that  $\hat{\omega}_{k+} \leq \hat{\omega}_{(k+1)+} \forall k \in \{1, 2, \dots, N\}$ .
- ii) *Select cut-off for subset selection when the rate of change of DTW distance is high:* calculate 'turning points' at index ' $k$ ' such that the absolute rate of change of DTW distance is highest in the local neighbourhood ( $\pm 1$  index),  $(\hat{\omega}_{(k-1)+} - \hat{\omega}_{(k-2)+}) \leq (\hat{\omega}_{k+} - \hat{\omega}_{(k-1)+}) \geq (\hat{\omega}_{(k+1)+} - \hat{\omega}_{k+})$ .
- iii) *Choose a turning point for subset selection:* choose the value at the first turning point ' $\hat{\omega}_k$ ' as our threshold  $\omega_{th}$  for finding the closest time series set  $\hat{\mathcal{X}}$ . The closest selected subset consists of all time series whose DTW distance is less than this threshold compared to the target time series.

Note that more turning points can be calculated by choosing the next minimum as described further.

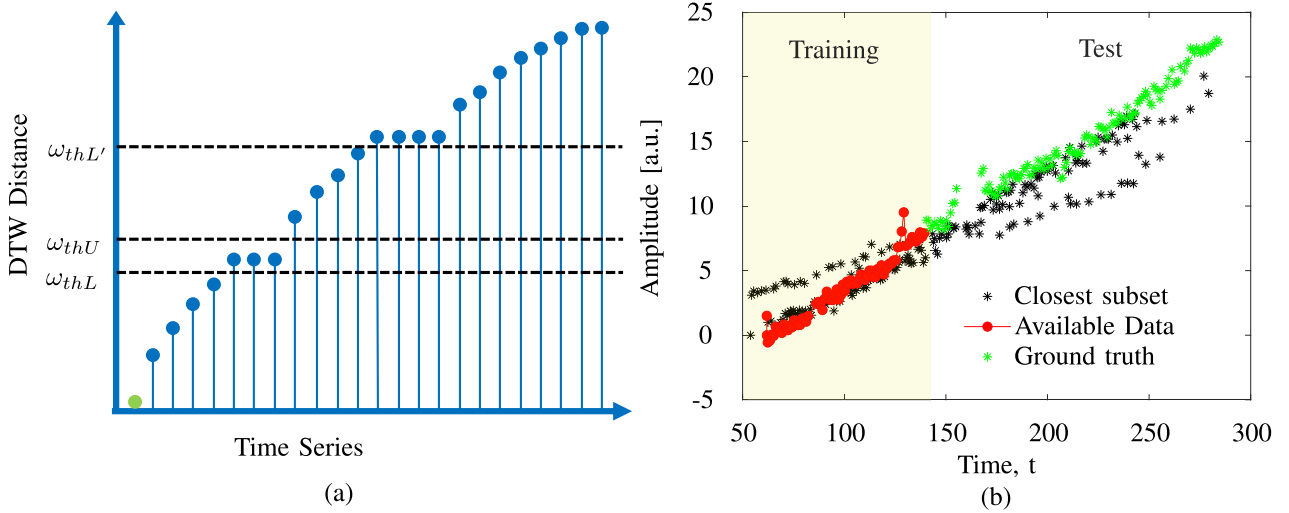


Fig. 4. (a) DTW distances, dissimilarity measures between time series, plotted in ascending order with some possible choices of threshold values. (b) Proposed heuristic is used to calculate the closest subset on the training part (in yellow,  $t < t_d^+$ ) and the test part for subject 1. This illustration is from the gestational weight gain prediction use-case explained in Section VI-B1.

The intuition for *turning points* is represented in Fig. 4(a), which shows the DTW distances measures between response variables of target and the training time series in ascending order. The possible choices of thresholds calculated as defined by *turning points* occur at locations  $\omega_{thL}$ ,  $\omega_{thU}$ ,  $\omega'_{thL}$ ,  $\dots$ ,  $\omega'''_{thL}$ . Note that  $\omega_{thL}$  represents the point where the first minimum occurs in the rate of change in DTW distances. Similarly, multiple such turning points exist that can be used as thresholds represented with the prime (') symbol. Intuitively,  $\omega_{thU}$  can be considered as another appropriate choice for threshold. However, the first value of turning point,  $\omega_{thL}$  is chosen as the preferred threshold. It selects the “smallest” most informative subset from the training data to capture the trend while keeping the variability among the selected subset to a minimum as compared to other thresholds. For the sake of simplicity, Fig. 4(b) shows a univariate time series of subject 1 from a dataset (explained in Section VI-B1) and the selected closest subset according to the proposed heuristics. Using the proposed heuristics, the subjects that are closer in the training phase (coloured in red) show a similar trend in the forecasting phase.

Using the SS approach proposed above, we can find a subset  $\hat{\mathcal{X}}$  from  $\{\mathbf{X}s, \mathbf{y}s\}$ . The subset  $\hat{\mathcal{X}}$  contains time series that are similar to the target time series and are therefore expected to contain the most essential information for forecasting the target time series data. The subset  $\hat{\mathcal{X}}$  will be used to train a non-parametric GP in the proposed SS-GP approach. The computational complexity of a GPs depends on the number of training points  $n$  according to  $O(n^3)$ . Restricting the training of the GPs to the subset  $\hat{\mathcal{X}}$  will considerably reduce the computational complexity (as compared to a training on the complete data set  $\mathbf{X}s$ ) because  $n(\hat{\mathcal{X}}) \ll n(\mathbf{X}s)$ . Moreover, we will show through our case studies that an increase in prediction performance can be obtained.

Additionally, such a localised non-parametric distance-based approach allows for the selection of neighbours based on the temporal nature of the data. This makes our approach generally

---

**Algorithm 1:** Temporal realignment for target data.

---

```

1: procedure TEMPORAL REALIGNMENT
2:   Input :  $\mathbf{y}^+ = [y^+(t_1^+) y^+(t_2^+) y^+(t_3^+) \dots y^+(t_d^+)]$ 
3:   lags =  $[-\tau_d, \dots, -\tau_1, 0, \tau_1, \tau_2, \dots, \tau_d]$ 
4:   Output :  $\tau_{optimal} N \times 1$ 
5:   for  $i = 1$  to  $N$  do
6:      $\mathbf{y}^i = [y^i(t_1^i) y^i(t_2^i) \dots y^i(t_d^i)]$ 
7:      $minDist = Inf$ 
8:     for  $iter = 1$  to  $2d + 1$  do
9:        $\tau = lags(iter)$ 
10:       $curDist = dist(\mathbf{y}^+, \mathbf{y}^i_{(t+\tau)})$ 
11:      if  $curDist < minDist$  then
12:         $\tau_{optimal}(i) = \tau$ 

```

---

applicable with other learning methods where priors are formed based on the available closest time series data.

### B. Collective Temporal Realignment

Typically, it is assumed that the time series in the training data are available from some fixed time  $t = t_0$ . However, in practical scenarios, the time series in the dataset may have different onsets and rates of progression.

Dynamic time warping (DTW) accounts for the similarity in amplitude among time series by calculating the distance between them. It realigns the two time series non-linearly, onto a common set of instants such that the sum of the Euclidean distances between the corresponding points, is smallest. We propose a time series alignment based on the shape of the response variable. We try to find a time instant  $\tau_{optimal}$  with respect to the target response series such that when the response time series in the training dataset are lagged/led by  $\tau_{optimal}$ , their shape most resembles that of the target’s response time-series. For a given target response variable ( $\mathbf{y}^+$ ), we realign the time series in  $\mathbf{X}s$  in time.

We hypothesise that readjusting the training data with respect to the target data will result in better subset selection. The approach is as follows,

- 1) Given target data observations of the response variable until time  $t_d$ , calculate distance from lagged/led versions of  $N$  time series in the training data using the metric  $\sqrt{\left(\sum_{n=1}^d y^+(t_n^+) - y^i(t_{n+\tau}^i)\right)^2}$
- 2) For the  $j^{th}$  time series in the training data, the value of  $\tau_{k+}$  that minimises the above metric is  $\tau_{optimal}(j)$
- 3) We then create lagged/led versions of predictor and response variables in the  $j^{th}$  training data using  $\tau_{optimal}(j)$  for the given target time series. This gives us the temporal fitted lagged/led version of  $\mathbf{X}_s^{\text{aligned}}$ .

Dynamic nearest neighbour selection is then applied to get  $\hat{\mathcal{X}}$ . After temporal realignment and dynamic subset selection based on the available target data ( $\mathbf{y}$ ) in the training and test dataset we apply Gaussian processes based prediction on  $\{\hat{\mathcal{X}}, \mathbf{X}^+\}$  as it is most resilient to the missing data in time series. We use the selected  $M \ll N$  time series that are in the closest subset of a given time series along with (8).

Mean Absolute Error (MAE) is used as the performance metric to evaluate the regression performance.

$$MAE = \frac{1}{N} \sum_{j=1}^N |y(t_{m_j}^j) - y_{pred}(t_{m_j}^j)|$$

## VI. EXPERIMENTS

We start by describing the setup of our experiments and the methods that we use to benchmark the proposed SS-GP approach. Furthermore, we give a detailed description of the use cases we will treat.

### A. Baseline

*Parametric:* We fit a 3rd order polynomial on the response variable varying with time. First, an MLE estimate is made on all the subjects in training data. These model estimates are used as prior distribution to calculate a maximum-a-posteriori estimate (explained in Section IV-B) to learn a final model. The response variable for a given test subject is then predicted using this final model at a given time instant. This is done in a leave-one-subject-out fashion so that each subject's data is estimated once. A 3rd order polynomial is used as it provides the least mean absolute error among other orders (1 to 5) of polynomials for both the data sets.

*ARIMA:* ARIMA has a limitation that it only works well with uniformly sampled data. This is difficult when data are missing. We fit an ARIMA( $p, d, q$ ) model on the response variable of the target data as follows i) linearly interpolating the data to make the data evenly-spaced in time, ii) tuning the hyperparameters [28] to find an optimal autoregressive order, degree of differencing, and moving average order by performing a grid search, iii) using the optimised hyperparameters over the training part to forecast at a time instant (given data until day  $t_d$ ). This is done for each test subject.

*LSTM:* We evaluate an LSTM-based regression network with 200 hidden units. The training is done using Adam's optimisation to minimise the mean absolute error [29].

*AR-GP:* AR-GP are trained to forecast the response variable using the input and response features until time  $t$ . For each subject, the missing observations are filled using the forward filling approach, where data from a previous observation are carried over to the following observation. When the training matrix is completed, the parameters of AR-GPs are learned by minimising the negative log-likelihood [20].

*AR-GP + MICE:* Multivariate imputation by chained equations (MICE) is an imputation strategy for matrix completion [30]. It works by iteratively building predictive models to fill each specified variable in the matrix. Each variable is imputed using other variables in the dataset and the iterations are run until convergence is met. AR-GP works by first forward filling the data to complete the matrix for training. We also use the state-of-the-art MICE approach to impute the data and then apply AR-GP to compare the performance.

We evaluate these methods on different univariate and multivariate real-life datasets in a leave-one-subject-out cross-validation scenario. A detailed explanation of how the proposed model is compared with baselines in different datasets is described as follows.

### B. Datasets

Health progression modelling requires longitudinal data from a person that can provide long-term predictions for disease status of an individual. Often, this data exists in the form of electronic health records or sequence readings collected over time. Current state-of-the-art methods such as deep learning methods provide accurate models of individuals' health status in case of Big Data sets where both the number of individuals and the number of individual measurements through time are large [31]. However, in the presence of limited training data (small  $N$  and  $t_d^+ \approx 0$ ), such as when early disease discovery is of utmost importance, such approaches produce sub-optimal results. Our framework for time series-prediction in the absence of missing or limited data can enhance health prediction capabilities. Hence, we select two datasets from real life presented as follows:

1) *Gestational Weight Gain:* One health demographic is managing gestational weight gain among women. Approximately 70% of pregnant women gain either too little or too much weight at the end of their pregnancy in accordance with the Institute of Medicine recommended guidelines [32]. Inappropriate weight gain during pregnancy has been associated with short- and long-term health complications to the mother and baby. Thus, early recognition of signs of weight gain during pregnancy is essential [22]. In this study, data were collected from diverse subjects in Europe where 80 women in their fifth week of pregnancy or later were recruited from midwife practices in Eindhoven, The Netherlands. The weight data were collected by a WiFi-connected scale, Withings WS30.<sup>2</sup> The dataset is described in Table I. Note that this is a case of univariate time series

<sup>2</sup>[Online]. Available: <https://www.withings.com/>

TABLE I

DATASET DESCRIPTION FOR UNIVARIATE GESTATIONAL WEIGHT GAIN DATA

Attribute	Mean $\pm$ Std (80 subjects)
Age (years)	31 $\pm$ 3.5
Height (meters)	1.69 $\pm$ 0.07
Pre-pregnancy weight (kgs)	69 $\pm$ 15
Pre-pregnancy BMI (kgs/m <sup>2</sup> )	24 $\pm$ 4
Delivery (days)	277 $\pm$ 10
Weight Gained (kgs)	13.7 $\pm$ 4.7
Number of weight gain samples	59.83 $\pm$ 41.02

data where only one variable (weight gain) is measured with respect to time. A mobile application allowed participants to log their weights weekly, and the weight data was sent to the cloud.

The participants provided an informed consent pre-data collection and the study was approved by the Internal Ethics Committee for Biomedical Experiments of the involved organisations (ICBE Reference number 2015-0079 respectively).

We model the weight (gain)  $y$  as a function of time,  $(t_1^j, t_2^j, \dots, t_m^j)$  using the proposed approach. We achieve this by first normalising measured weight with pre-pregnancy weight to obtain weight gain data and then fitting various forecasting approaches. For the parametric approaches, we utilise the complete data from  $N - 1$  subjects to generate a prior to estimate a MAP model. We experiment with first, second and third-order polynomial based parametric approaches to fit our time-series data. In cross-validation, we obtain the polynomial order ( $= 3$ ) empirically for the parametric approach, which has the lowest prediction error among all other orders.

For the proposed non-parametric approach, we use the data from  $N - 1$  subjects as training data in addition to the available target data of the remaining subject to train the Gaussian processes in the baseline setting. Dynamic subset selection is performed on the training data with respect to the available target data.

**2) Alzheimer's Disease Prediction:** Another health complication is Alzheimer's disease (AD). AD is a neurodegenerative disorder and the most common form of dementia. Prediction of this progressive disorder's symptom onset at early stages is urgent and complex [33]. The design of clinical trials and developing therapeutic interventions depends on accurately detecting patients at the early stages of the disease where treatments are most likely to be effective. The clinical status of an Alzheimer's patient is based on commonly used cognitive scores namely, the mini mental state examination (MMSE) [34], the Washington University Clinical Dementia Rating Sum of Boxes score (CDRSB) [35], and the AD Assessment Scale-Cognitive subtest (ADAS-Cog13) [36].

To this end, we use the data collected as part of the TADPOLE challenge [37] by the Alzheimer's Disease Neuroimaging Initiative (ADNI) consortium<sup>3</sup> [38]. The data from 1737 patients

taken every six months over the course of 120 months consists of different modalities such as (1) various features extracted from imaging modalities like magnetic resource imaging (MRI), positron emission tomography (PET) and diffusion tensor imaging (DTI), (2) cerebro-spinal fluid (CSF) markers of amyloid beta and tau-deposition; (3) cognitive assessments measured in the presence of a clinical expert; (4) genetic information such as alipoprotein E4 (APOE4) status from DNA samples and (5) general demographic information [37]. Around 266 features were extracted based on these modalities and merged together over time to form a coherent numerical multivariate time series feature set. Since the complete dataset has a lot of missing visits, we follow the state-of-the-art approach for Alzheimer's disease marker forecast [20] and selected a smaller dataset of 95 subjects such that data from at least ten visits is present and missing data is no more than 82.5% of the feature set. This helps in benchmarking our proposed approach with the AR-GP approach [20].

In the case of this multivariate time series data, our experimentation to predict a cognitive score (MMSE, ADAS or CDRSB) using 266 features that vary with time is as follows:

- 1) Collective temporal realignment: The Alzheimer's study [38] recruited patients that were already going through some stage of cognitive decline. Since the disease progression in every individual differs in their onset, the target time series ( $y$ ) for each of the patients had a different  $t_0$ . Therefore, we calculate the value of  $\tau_{optimal}(j)$  using the response variable of the target data and the response variable of the  $j^{th}$  subject in the training data. This lag is calculated for all the subjects in the training data with respect to a given target subject. Based on the calculated  $\tau_{optimal}(j)$ , lagged/led versions of the predictor ( $X^j$ ) and the response ( $y^j$ ) are created to be used for further training.
- 2) Subset selection based on the response variable  $y$ : Based on the available target data ( $X, y$ ) from a given test subject until month  $t_d$ , we find the subjects in the training dataset with a similar cognitive decline. This is done by applying the subset selection approach explained in Section V-A on  $y^j$ . Note that we apply subset selection on the response variable instead of  $X$  since it gives us similar subjects in output space.

We also compared the performance of our subset selection approach with a k-means clustering approach. Clusters were obtained using the input features of the multivariate time series. For a given test subject, the subjects in the closest cluster are considered as training data. These training subjects along with the available test subject's data are used to train a non-parametric GP as follows: (1) the missing values are forward filled (2)  $K$  clusters with centroids  $\{c_k\}$  are created using k-means clustering with training data matrix until month  $t_d$  (3) calculate distance  $d_k = dist(X^+, c_k)$  of the test subject's predictor variables data  $X^+$  from each centroid  $\{c_k\}$  and the optimal profile (subset) is selected as cluster  $c_{opt} = c_i$  such that,  $d_i < d_j, \forall i \neq j \in [1, K]$  In what follows, we will refer to this method as the "K-means + GP" approach.

<sup>3</sup>[Online]. Available: <http://adni.loni.usc.edu/>



**Disease progression Estimation:** We perform non-parametric regression using Gaussian processes on the input feature set  $\mathbf{X}$ . First, given a test subject's response variable  $\mathbf{y}^+$ , a time aligned training data is made that consists of a lagged version of  $\mathbf{X}$ s and  $\mathbf{y}$ s. Subsequently, subset selection is performed by finding subjects in the training data whose response variable ( $\mathbf{y}^j$ ) is close to the test subject's response variable ( $\mathbf{y}^+$ ). Once a subset is selected, GP based regression is performed on  $[\mathcal{X}, \mathbf{X}^+]$  using (7) and (8). We perform leave-one-subject out cross-validation on the dataset.

In both cases, developing models to automatically predict Alzheimer disease-related metrics or gestational weight gain is of utmost importance to intervene appropriately and in time. This makes the availability of the target data another challenge. To test how well these methods can perform with limited target data, we experiment by varying the amount of available target data with respect to time, i.e.,  $0 \leq t_d^+ \leq t_m^+$ .

Remark that, in the case of gestational weight gain prediction, the objective is to predict a single observation in time, i.e., the end-of-pregnancy weight gain. The performance is measured by predicting the end-of-pregnancy ( $\approx 270$  day) weight gain for a test subject when data was available until 120, 130, 140,  $\dots$ , 260 days.

In the case of Alzheimer's disease, however, we are also interested in the disease's trajectory, not merely an ultimate endpoint prediction. **Two subsequent visits are spaced an average of 6 months apart, and we will predict the disease progression for each month despite having little data** (i.e., only observations from the first 30 months are used in the training phase to predict progression up to month 120).

We evaluate and present the results related to the performance of different approaches across time with different availabilities of the target data. We also benchmark our proposed approach with the K-means + GP approach and the AR-GP approach [20], the latter of which is considered as a state-of-the-art approach for predicting cognitive decline of Alzheimer's patients.

## VII. RESULTS & DISCUSSION

### A. Gestational Weight Gain Prediction

We study the performance of various forecasting algorithms when predicting the weight gain of a target subject for the end of the pregnancy while the time series data of weight gain of the target subject are only available up to time  $t_d^+$ . The most crucial aspect of gestational weight gain prediction is whether the weight at the end of the pregnancy is in the range recommended by the IOM guidelines [32]. Therefore, we investigate the weight gain prediction ability of forecasting algorithms that are trained with changing availability of a target individual's data (i.e., varying  $t_d^+$ ). The results can be found in Fig. 5, which shows the prediction error averaged over all the subjects as a function of the moment  $t_d^+$ . The prediction error reduces when more training data is available. Also, it can be observed in Fig. 5 that the GP approach performs worse than the SS-GP approach. Based on a paired t-test, which assumes equal variances, we found that all differences between performances of the SS-GP model and the other models are statistically significant at a

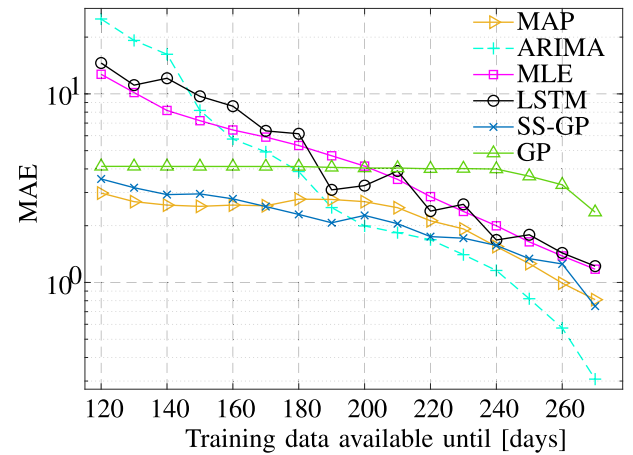


Fig. 5. MAE of predicted weight on delivery day (multiple steps ahead in time) with respect to different approaches. MAE reduces as more training data becomes available.

significance level of 5%. Only for the SS-GP and the MAP model performances, no statistically significant difference was found. This is not unexpected because of the simplicity of the dataset. For  $t_d^+ > 220$ , the performance of ARIMA significantly outperforms the performance of all other approaches. However  $t_d^+ = 220$  is too close to the horizon to result in effective intervention. Note that the average delivery day is around day 277. The benefit of using our SS-GP approach is further illustrated in Fig. 6.

In Fig. 6(a), we show the performance of GPs when all training data is used to make predictions for a target subject. Since the training data consists of subjects with various rates of weight gain, the predicted trend in the target subject is influenced by all the measurements in the training data at a given time. The variability in the prediction is reduced by selecting a subset of time series from the training data that share similar patterns with the target data. These subjects are then used to forecast the target data, as shown in Fig. 6(b).

### B. Alzheimer's Disease Prediction

Unlike the gestational weight gain use case, where the final objective was to predict the end-of-pregnancy weight gain because the data was recorded daily, we aim to predict the progression of Alzheimer's disease at each visit, since these visits are separated by six months or more. For this purpose, we will study the performance of several methods for predicting three metrics for cognitive decline that are commonly used by clinicians and that were introduced in Section VI-B2: MMSE, ADAS13, and CDRSB.

Following our realignment approach, we first calculate the optimal  $\tau$  for each response time series (cognitive score) in the training dataset with respect to the available response data from the test subject. We compute the standard deviation at a particular time instant for all response time series in the training data that are aligned with respect to the target subject. This standard deviation should be smaller than when no alignment is performed. We experimented with all the subjects in a leave one

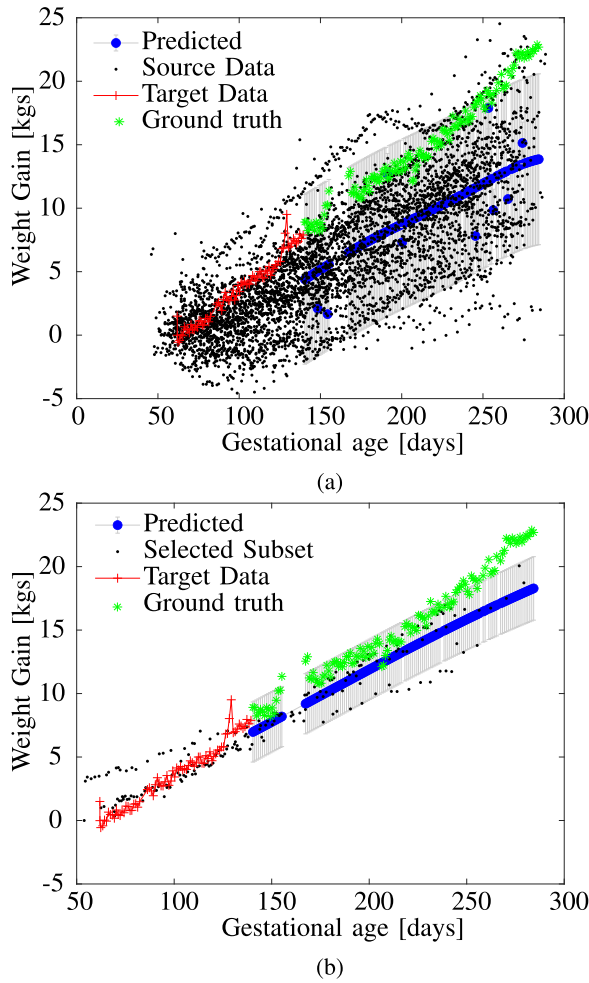


Fig. 6. Prediction error ( $i = 1^{th}$  subject) is (a) high (low confidence) when the complete training dataset is considered due to inter-subject differences but (b) reduces using close subset selection based on heuristics. The prediction confidence (grey) also increases using the SS approach.

out fashion. In Fig. 7 each line depicts the standard deviation of the ADAS13 matrix created using aligned versions of the time series for a given test subject. By computing the standard deviation without alignment, a baseline was established. We observed that  $>80\%$  of the subjects have a standard deviation less (more desirable) than the baseline when adjusted for the alignment using our temporal realignment approach. This shows that most of the subjects are adjusted in time with respect to disease progression after realignment.

To predict the Alzheimer's disease progression, we varied the availability of target data from month 30 until month 108. Fig. 8 shows the cross-validation results, averaged over all subjects, for the prediction of ADAS13 using our SS-GP approach. Each line in Fig. 8 corresponds to a different number of available measurements for the test subject. Given the available training data until a specific month, each point on this line represents the prediction error in forecasting cognitive score for the month depicted on the x-axis, averaged over all subjects. One can observe from Fig. 8 that there is an increasing trend in prediction error when the forecast horizon increases. For example, given

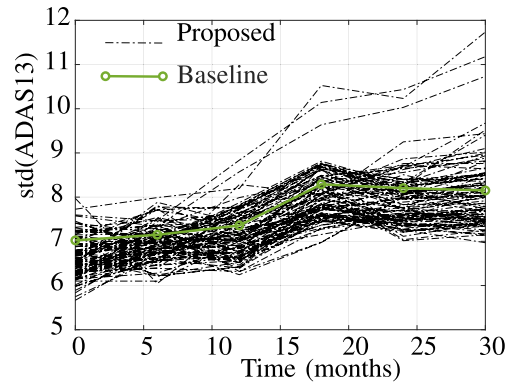


Fig. 7. Standard deviation (std) of the cognitive decline (ADAS13) after the proposed alignment for each subject (in black). The closer the std is to the x-axis, the more similar the subjects' time series are.

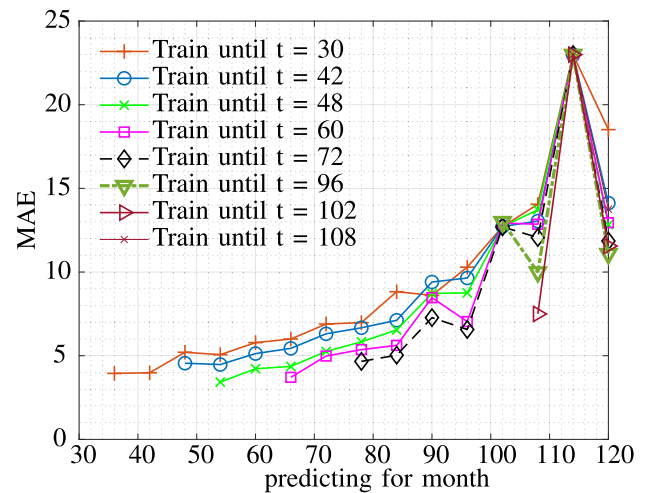


Fig. 8. Mean absolute error measured with respect to different data available in time for different steps in time prediction for ADAS13. The average MAE for a specific month is lower when the data availability is higher.

the training data availability until month 30 (orange line with + marker), the mean absolute error when predicting for month 60 is higher than for month 36. Additionally, Fig. 8 shows that the prediction performance improves as more data from the test subject becomes available for training. For instance, for the predictions at month 48, the MAE obtained when training data are available up till month 42 is smaller than the MAE obtained when training data are available up till month 30. The other metrics for cognitive decline (MMSE and CDRSB) were found to show similar patterns in prediction performance.

As seen in Fig. 8, the worst result is observed when the available training data is highly limited, and the forecast horizon is set far in time. This occurs when training data are only available until month 30 and forecasts are made up till month 120. Since we need to predict as early as possible, we present the results for all further experiments when training data are only available until month 30, and the forecast horizon stretches from month 36 to month 120. Using training data up till month 30 ensures us that during training at least one data point of each subject is included.

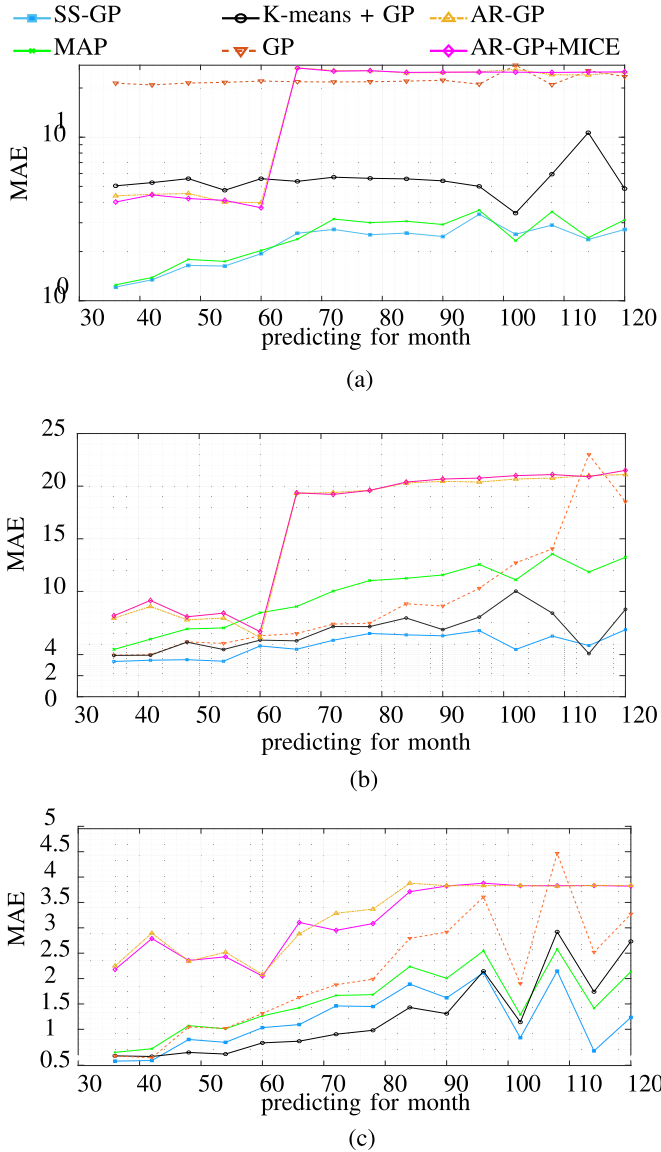


Fig. 9. Proposed approach achieves lowest MAE on the metrics (a) MMSE and (b) ADAS13 and comparable MAE with k-means based clustering on (c) CDRSB.

Furthermore, at month 114, a peak in MAE is observed in Fig. 8. This is due to the fact that, at month 114, no measurements of the target variable are available for a lot of subjects.

Fig. 9 shows the forecast performance for the three cognitive metrics for different subset selection strategies and different regression approaches. For the AR-GP approach a forward filling approach is used to deal with missing data [20]. However, we also studied the performance of the AR-GP approach when a state-of-the-art imputation technique is used instead, i.e. a multivariate imputation by chained equations (MICE) for matrix completion [30]. We refer to this combination as AR-GP + MICE.

Note that on average, a 1-3 point decrease in Mini Mental State Examination [39], a 1-2 point increase in Clinical Dementia Scale sum of boxes [39], and a 3-3.1 point increase in Alzheimer's Disease Assessment Scale-Cognitive (ADAS-Cog) [40] are indicative of a meaningful decline.

The differences between the proposed and compared models are statistically significant ( $p < 0.05$ ) based on a paired t-test with equal variances. However, compared to the SS-GP approach, no statistical difference is found with the MAP approach when predicting MMSE and with the K-means + GP approach when predicting ADAS13 or CDRSB. Thus, we can conclude that our method performs consistently (equal if not better) across all metrics of cognitive decline when compared with the state-of-the-arts.

## VIII. CONCLUSION

In this article, we proposed a novel approach, termed the SS-GP approach, for forecasting time series that are not necessarily uniformly sampled. For this purpose, we combined the non-parametric GP regression with a subset selection procedure that selects a set of time series from the data that closely resembles the test subject's data. Our subset selection procedure is robust as it selects the subset size dynamically based on temporal similarities between the time series in the subset and the test time series. The temporal similarity is measured with a DTW distance that can be computed between time series with a different length. We validated this method on two use cases and compared it with several other approaches.

Firstly, on the univariate gestational weight gain dataset, our approach performs similar to a parametric polynomial fitting which is not unexpected because of the simplicity of the data set. However, the SS-GP is able to reduce the variability in predictions because predictions are only based on time series data that share similar patterns with the data of the test subject.

Secondly, for a more complex data set consisting of multivariate time series data to predict cognitive decline of Alzheimer's patients our SS-GP approach is able to outperform state-of-the-art approaches such as the AR-GP approach [20]. In particular, the SS-GP approach, improves prediction results when the forecast horizon is long and only a limited amount of data is available.

## IX. LIMITATIONS & FUTURE WORK

Although effective in regression when data is missing, Gaussian Processes (GPs) have a high computational complexity of  $\mathcal{O}(n^3)$ . Our subset selection is a local approximation technique that decreases complexity by including only the most useful training points ( $\ll n$ ) that are close to the test point. However, the collective realignment technique has a high time complexity because it determines the ideal alignment for a specific test time series by comparing it to all the time series in the training dataset. In future research, we would like to experiment with another scalable sparse approximation of GPs developed in [41] that can further reduce the time complexity.

In addition, the proposed approach is only tested on data sets from healthcare considering the necessity that arises in this domain from data acquisition limitations. In an environment where data acquisition is costly, it would be beneficial to evaluate our method on more data sets and application domains where time series can be sparsely sampled, such as process quality monitoring in industries.



## ACKNOWLEDGEMENT

This publication reflects only the authors' view and the REA is not responsible for any use that may be made of the information it contains.

## REFERENCES

- [1] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications: With R Examples*. New York, NY, USA: Springer, 2017.
- [2] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhelou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31314–31338, 2015.
- [3] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel, "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models," in *Proc. 2nd ACM SIGHIT Int. Health Inform. Symp.*, 2012, pp. 389–398.
- [4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computat.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [5] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long-and short-term temporal patterns with deep neural networks," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2018, pp. 95–104.
- [6] N. Laptev, J. Yosinski, L. E. Li, and S. Smyl, "Time-series extreme event forecasting with neural networks at uber," in *Proc. Int. Conf. Mach. Learn.*, vol. 34, 2017, pp. 1–5.
- [7] E. De Brouwer, J. Simm, A. Arany, and Y. Moreau, "GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 7379–7390.
- [8] M. Liu, A. Zeng, Z. Xu, Q. Lai, and Q. Xu, "Time series is a special sequence: Forecasting with sample convolution and interaction," 2021, *arXiv:2106.09305*.
- [9] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced Lectures on Machine Learning*. Berlin, Germany: Springer, 2004, pp. 63–71.
- [10] J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting," *Int. J. Forecasting*, vol. 22, no. 3, pp. 443–473, 2006.
- [11] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. Neural Inf. Process. Syst. Workshop Deep Learn.*, 2014.
- [12] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2015.
- [13] K. Cho et al., "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.
- [14] Z. C. Lipton, D. Kale, and R. Wetzel, "Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series," in *Proc. Mach. Learn. Healthcare Conf.*, 2016, pp. 253–270.
- [15] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1519–1528, May 2021.
- [16] K. Li, C. Liu, T. Zhu, P. Herrero, and P. Georgiou, "GluNet: A deep learning framework for accurate glucose forecasting," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 414–423, Feb. 2020.
- [17] J. Futoma, S. Hariharan, and K. Heller, "Learning to detect sepsis with a multitask Gaussian process RNN classifier," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1174–1182.
- [18] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PLoS One*, vol. 13, no. 3, 2018, Art. no. e0194889.
- [19] L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian process regression in vital-sign early warning systems," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2012, pp. 6161–6164.
- [20] K. Peterson, O. Rudovic, R. Guerrero, and R. W. Picard, "Personalized gaussian processes for future prediction of alzheimer's disease progression," in *Proc. Neural Inf. Process. Syst. Workshop Mach. Learn. Healthcare.*, 2017.
- [21] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances," *Data mining Knowl. Discov.*, vol. 31, no. 3, pp. 606–660, 2017.
- [22] C. Puri et al., "PREgDICT: Early prediction of gestational weight gain for pregnancy care," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2019, pp. 4274–4278.
- [23] E. J. Keogh and M. J. Pazzani, "A simple dimensionality reduction technique for fast similarity search in large time series databases," in *Proc. Pacific-Asia Conf. Knowl. Discov. Data Mining*, Springer, 2000, pp. 122–133.
- [24] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. KDD Workshop*, Seattle, WA, 1994, vol. 10, no. 16, pp. 359–370.
- [25] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowl. Inf. Syst.*, vol. 7, no. 3, pp. 358–386, 2005.
- [26] M. Müller, "Dynamic time warping," in *Information Retrieval for Music and Motion*. Berlin, Germany: Springer, 2007, pp. 69–84.
- [27] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, "Generalizing DTW to the multi-dimensional case requires an adaptive approach," *Data Mining Knowl. Discov.*, vol. 31, no. 1, pp. 1–31, 2017.
- [28] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. 63, no. 1, pp. 117–126, 1976.
- [29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
- [30] S. v. Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, pp. 1–68, 2010.
- [31] S. El-Sappagh, T. Abuhmed, S. R. Islam, and K. S. Kwak, "Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data," *Neurocomputing*, vol. 412, pp. 197–215, 2020.
- [32] K. M. Rasmussen, P. M. Catalano, and A. L. Yaktine, "New guidelines for weight gain during pregnancy: What obstetrician/gynecologists should know," *Curr. Opin. Obstet. Gynecol.*, vol. 21, no. 6, 2009, Art. no. 521.
- [33] J. L. Cummings, "Challenges to demonstrating disease-modifying effects in Alzheimer's disease clinical trials," *Alzheimer's Dement.*, vol. 2, no. 4, pp. 263–271, 2006.
- [34] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "'Mini-mental state': A practical method for grading the cognitive state of patients for the clinician," *J. Psychiatr. Res.*, vol. 12, no. 3, pp. 189–198, 1975.
- [35] C. P. Hughes, L. Berg, W. Danziger, L. A. Cohen, and R. L. Martin, "A new clinical scale for the staging of dementia," *Brit. J. Psychiatry*, vol. 140, no. 6, pp. 566–572, 1982.
- [36] W. G. Rosen, R. C. Mohs, and K. L. Davis, "A new rating scale for Alzheimer's disease," *Amer. J. Psychiatry*, vol. 141, pp. 1356–1364, 1984.
- [37] R. V. Marinescu et al., "TADPOLE challenge: Accurate Alzheimer's disease prediction through crowdsourced forecasting of future data," in *Proc. Int. Workshop Predict. Intell. Med.*, Oct. 13, 2019, pp. 1–10.
- [38] S. G. Mueller et al., "The Alzheimer's disease neuroimaging initiative," *Neuroimaging Clin.*, vol. 15, no. 4, pp. 869–877, 2005.
- [39] J. S. Andrews, U. Desai, N. Y. Kirson, M. L. Zichlin, D. E. Ball, and B. R. Matthews, "Disease severity and minimal clinically important differences in clinical outcome assessments for Alzheimer's disease clinical trials," *Alzheimer's Dementia: Transl. Res. Clin. Interv.*, vol. 5, pp. 354–363, 2019.
- [40] A. Schrag et al., "What is the clinically relevant change on the ADAS-Cog?," *J. Neurol., Neurosurgery Psychiatry*, vol. 83, no. 2, pp. 171–173, 2012.
- [41] E. Snelson and Z. Ghahramani, "Local and global sparse Gaussian process approximations," in *Proc. Artif. Intell. Statist.*, 2007, pp. 524–531.