Exploring Missing Data Prediction in Medical Monitoring: A Performance Analysis Approach

Qiong Gui, Zhanpeng Jin
Department of Electrical and Computer Engineering
Binghamton University, State University of New York (SUNY)
Binghamton, NY 13902-6000
Email: {qgui1, zjin}@binghamton.edu

Wenyao Xu
Department of Computer Science and Engineering
University at Buffalo, State University of New York (SUNY)
Buffalo, NY 14260-2500
Email: wenyaoxu@binghamton.edu

***Abstract*—Medical monitoring represents one of the most critical components in existing healthcare system. The accurate and reliable acquisition of various physiological data can help physicians and patients to properly detect and identify potential health risks. However, this process suffers from severe limitations in terms of missing or degraded data, which may lead to a rather high false alarm rate and potentially compromised diagnostic results. In this paper, we investigated three different approaches for missing data prediction in clinical settings: mean imputation, Gaussian Process Regression (GPR), and Kalman Filter (KF). Experimental results show that, the heart rate (HR) signals largely rely on most recent data and missing data prediction will be less effective for further prediction.**

I. INTRODUCTION

Health is a critical component of human well-being and has raised increasing concerns from the entire world. It was estimated by the Department of Health and Human Services that the health share of GDP will continue its steady growth [1]. Similar results were reported by the Organization for Economic Co-operation and Development (OECD) [2]. The great challenges for healthcare are the increased life expectancy and the consequently increased aging population. A vital and costly part of the current healthcare system is the monitoring of patients' relevant physiological signals, such as heart rate (HR), respiratory rate (RR), blood pressure (BP) and oxygen saturation level, to accurately identify and indicate the health status of each individual. Driven by the huge demands of healthcare services and caregiving workloads, the healthcare system is now transforming from a physician-centric and hospital-centric model to patient-centric, continuously monitoring manner [3]. In this sense, the quality of healthcare heavily rely on the quality of those medical monitoring systems.

However, it is not always the case that the monitor can obtain the physiological signals flawlessly. It is not uncommon that, the acquired physiological signals are strongly impaired by noises; the sensors are temporarily or permanently disconnected from the patient (due to the patient's movement or loose adhesion) and thus cannot detect anything; there may be errors resulted from transmission or recording, or omitted by human. Some pieces of physiological signals can also be simply ignored since the data is less important and less used. Or the records may not match with each other because of asynchronous clocks, and so on [4]. Unfortunately, for traditional patient monitors, they generate alarms if any of monitored physiological parameters exceeds a certain range, which becomes problematic for the case when the data is missing. In this case, a huge amount of false alarms may be generated just because of missing or corrupted data. Table I shows the missing rate of some patients' heart rate and blood pressure records from the MIMIC II database [5]. The missing rate varies a lot for different patient records, from 0.19% to an extraordinary high level of 38.95%. Those missing data can cause rather high level of false alarms which can not only increase the workload of the caregivers, causing them fatigue [6], but also increase the life threatening risk of the patients [7]. Several studies have reported an extremely high rate of false alarms in critical care monitoring (up to 90% of all alarms are false positives) while the vast majority of such alarms have no real clinical impact [8]–[10]. Since the missing data can decrease the accuracy of decision making, it is important to minimize its influence. If we can make a reasonable prediction that is very close to its actual value, this would be very helpful. As it is well demonstrated in other literature that the physiological signals show strong spatial-temporal relations, it is possible that we can predict the missed pieces based on historical data.

Missing data problem has been extensively studied in a wide variety of domains, such as economics, finance, and engineering. Some existing methods can be properly applied in patient monitoring. Acute Physiology, Age and Chronic Health Evaluation (APACHE) system [11] is an accepted scoring system derived from physiological signals, blood tests and arterial blood gas tests. It assumed the missing data was in normal state and used the mean value to represent the part. In [12], the authors presented a weighted K-nearest neighbor algorithm on multi-dimensional intensive care unit (ICU) clinical data. The signal quality index method considered the effect of the signal quality [13] and used other more reliable physiological signals to extract the clinical features. But this method required other different sources of signals.

In this paper, we present a comparative analysis of the performance of mean imputation, Gaussian Process Regression (GPR), and Kalman Filter (KF) approaches. Since HR and BP are vital signs having similar characteristic in the missing data prediction, here we use the HR as an example to investigate the performance of the three different methods. In the following sections, we first describe these three algorithms. Then based on the described experimental setting, we evaluate the predicted results on the HR from MIMIC II database and analyze how the window size and prediction step influence the results.

TABLE I. MISSING RATE IN MIMIC II DATABASE

Record	Total	Heart Rate	%	Blood Pressure	%
a40017	3610	571	15.82	1406	38.95
a40022	8633	192	2.22	5960	6.90
a40076	4228	8	0.19	212	5.01
a40084	3887	102	2.62	396	10.19
a40093	2677	372	13.90	401	14.98
a40154	4090	67	1.64	249	6.09
a40414	4663	382	8.19	476	1.02
a40471	1668	40	2.40	74	4.44
a40502	11117	115	1.03	1183	10.64
a40645	6749	287	4.25	465	6.89
Total	51322	2136	4.16	10822	21.09

II. ALGORITHMS

As many algorithms are developed to predict the missing data in various areas, they can be generally classified into four categories: completely recorded units based, imputation based, weighting procedure based, and model based methods [14]. The completely recorded units based method simply discards the incompletely recorded units and only uses the complete data for analysis purpose. The imputation-based procedure uses the standard methods to analyze the resultant completed data after the missing values are filled in. The weighting procedure analyzes the data by their contribution. The predict values will rely on the more relevant points. The model-based procedure builds a model based on the partially missing data and uses this model to predict the missed values. Since this method is on the likelihood, it is more flexible. Specifically, mean imputation, GPR and KF are among the most representative algorithms for prediction.

A. Mean Imputation

In mean imputation, it simply uses the average value of measured data to replace the missed ones.

B. Gaussian Process Regression

Gaussian Processes (GPs) [15] extend multivariate Gaussian distributions to infinite dimensionality. There are two functions involved in GPR: the mean function which is assumed to be 0 everywhere, and the covariance function:

$$k(x,x') = \sigma_f^2 exp\left[\frac{-(x-x')^2}{2l^2}\right] + \sigma_n^2\delta(x,x') \qquad (1)$$

where σ_f^2 is the signal variance, l is the length scale, σ_n^2 is the noise variance, $\delta(x,x')$ is the Kronecker delta function. Based on the covariance, we can get the best estimation by the format of mean of the distribution and the uncertainty in variance. A sparse pseudo-input Gaussian process [16], [17] was proposed to improve the computational efficiency and relaxing the restrictive modeling assumptions of the standard Gaussian process by sparse approximation based on a set of M "pseudo-inputs". For these advantages, we used the sparse pseudo-input method for GPR prediction.

C. Kalman Filter

Kalman Filter is a state space model based optimal estimator. It can be expressed in the following way:

$$\begin{array}{ll} x(k+1) & = Ax(k) + Bu(k) + w(k)(\text{state}) \\ y(k) & = Cx(k) + v(k)(\text{measurement}) \end{array} \qquad (2)$$

where, $x(k)$ is the state, $y(k)$ is the observation, $u(k)$ is the outside input, $w(k)$ and $v(k)$ are noises and have zero-mean and normal probability distribution, A is the state transition model applied to the previous state, B is the control-input model applied to the outside input, C is the observation model. Based on the above two equations, one gets the prediction by minimizing $E\left[(x_k - \hat{x}_k)(x_k - \hat{x}_k)^T\right]$. To predict the patients' vital signs, like HR, the state equation can be simplified by setting $A = 1$, $B = 0$, and $C = 1$, let the HR be both the measurement and the state describing the process [18], [19].

Generally, KF relies on the current observation and requires less memory to store previous information. For 1-step prediction, the predicted value is calculated as following equation:

$$\hat{x}\left[N+1|N\right] = A\hat{x}\left[N|N-1\right] + K\left(N\right)e\left(N\right) \qquad (3)$$

When two or more points ahead need to be predicted, but the new observations are unknown, the r-step ahead prediction value is calculated as the following equation using the data we currently known:

$$\hat{x}\left[N+r|N\right] = A^{(r-1)}\hat{x}\left[N+1|N\right] \qquad (4)$$

The multi-step predictions mainly based on the state transition model A and the prediction depth r.

III. RESULTS

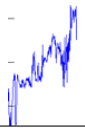
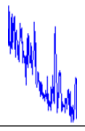
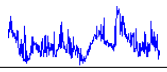
A. Experiment Setting

The MIMIC II database [5] was used as a source of data for the experiment and the numerical heart rate data which was one sample per minute was selected as the testing data. The records of 10 patients were selected. Each record of the selected 10 patients was first separated into smaller segments without unknown or zero data by the trends which were shown in Table II. Since when we kept receiving data, we knew the trend in the short time period and the heart rate seemed to hold the short time trend. By analyzing the performance of each trend, we could evaluated how reliable we could trust on the predictions.

Given the fact that the window size of the historical information may influence the results, we would like to investigate a more effective way to predict more data because it is very often that two or more consecutive data samples are missing. In this study, we randomly chose a point, shown by the red arrow in Figure 1, to decide where to start the prediction. To evaluate the influence of the window size, we included more or less data from the red point to the left. To analyze how many points we could predict that were close to the actual values, 10 points data on the right of the red point from 1 to 20 with incremental of 2 were predicted. Since the multi-step prediction was actually using only the data already known, we thus used the fixed window without sliding. After we got the predictions, we compared them with the original data using the Mean Squared Error (MSE) in the following equation to evaluate the performance:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}_i - Y_i\right)^2 \qquad (5)$$

TABLE II. DESCRIPTION OF TESTING DATASETS

Trend	Screen Shot	Description	# of sets
STABLE		almost a constant value	21
INCREASE 1		stably increasing segment	15
DECREASE 1		stably decreasing segment	16
INCREASE 2		general trend is increasing, but data vary a lot for a short period of time	22
DECREASE 2		general trend is decreasing, but the data vary a lot for a short period of time	19
OBSCURE		data change frequently in large amplitudes	53

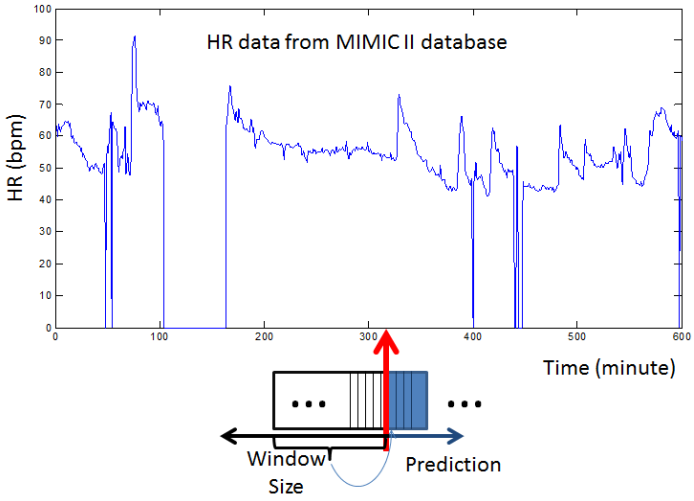


Fig. 1. Experiment Window Size and Prediction Length Set Up

where N is the total number of predictions, Y_i is the original value, $\hat{Y}_i$ is the predicted values.

B. Experiment Results

In order to better compare the performance for different window sizes and prediction lengths, we plotted two figures for each approach. One showed the MSE distribution of different window sizes from 1 to 56 with incremental of 5 on 1-step, 2-step and 3-step prediction. The other showed the MSE distribution of different prediction lengths on fixed window sizes of 5, 10 and 15. The box plot was chosen to show the spread degree of the results. The white circle with black dot inside represented the median of the test results. The edges of the box were the 25th and 75th percentiles. The vertical

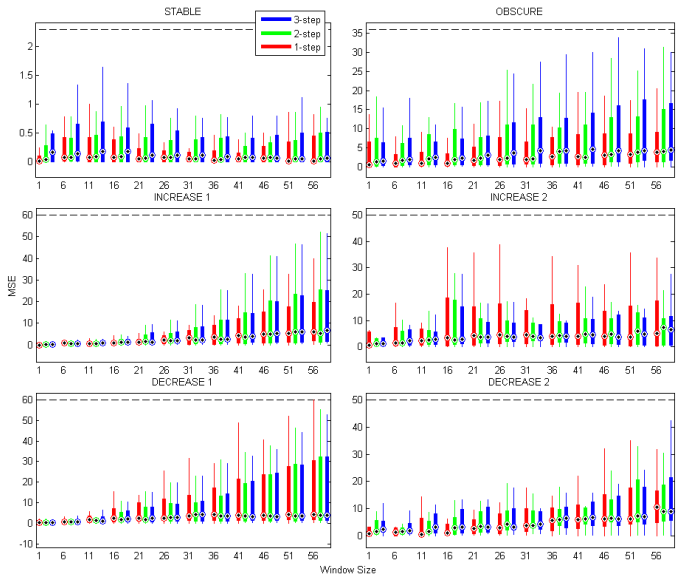


Fig. 2. Results of Mean Imputation Based Prediction (Different Window Sizes)

lines indicated the variability outside the two quartiles up to the extreme data.

Figure 2 presents the MSE values using mean imputation over different window sizes and the color boxes represented the results of different prediction lengths. From this figure, we can see that the STABLE trend had the best performance. The predicted values were very close to their expected values. The maximum error was no larger than 1 for 1-step and 2-step prediction and 2 for 3-step prediction. Both INCREASE 1 and DECREASE 1 cases had the results that the errors were small when the window size was not large for all the three different step predictions. When the window size was no greater than 16 for INCREASE 1 and 11 for DECREASE 1, the MSEs were less than 5. When the window size exceeded these values, the errors would increase rapidly. The errors in INCREASE 2 and DECREASE 2 were larger. Although a few MSEs can reach over 10 when the window size was small, most of the MSEs were distributed in the range less than 10 when the window size was no larger than 16 for INCREASE 2 and 41 for DECREASE 2. For INCREASE 2, the errors range at window size 16 extended to around 20. As the window size increased, the MSE range would drop or rise slowly. Compared to 2-step and 3-step prediction, 1-step prediction had the worst results. For the OBSCURE case, when the window size was no larger than 26, most of the errors were in the range less than 5 for prediction length of 1 while 2-step and 3-step prediction had most MSEs less than 10. Since the blue boxes were higher than green ones while the green boxes were higher than the red boxes in general when the window size was larger and equal than 26, the accuracy of predicting more data decreased.

Figure 3 shows the results of the MSEs for different prediction lengths and the color boxes show the results on different window sizes of 5, 10, and 15 respectively. The predictions were very accurate at the STABLE trend no matter how many data we wanted to predict because the maximum error was no larger than 3.5. The MSEs were increasing for both INCREASE 1 and DECREASE 1 as more data were

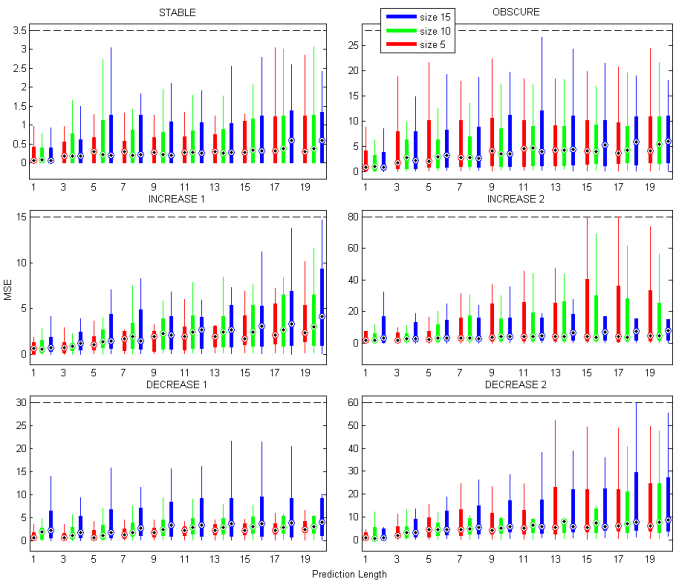


Fig. 3. Results of Mean Imputation Based Prediction (Different Prediction Lengths)

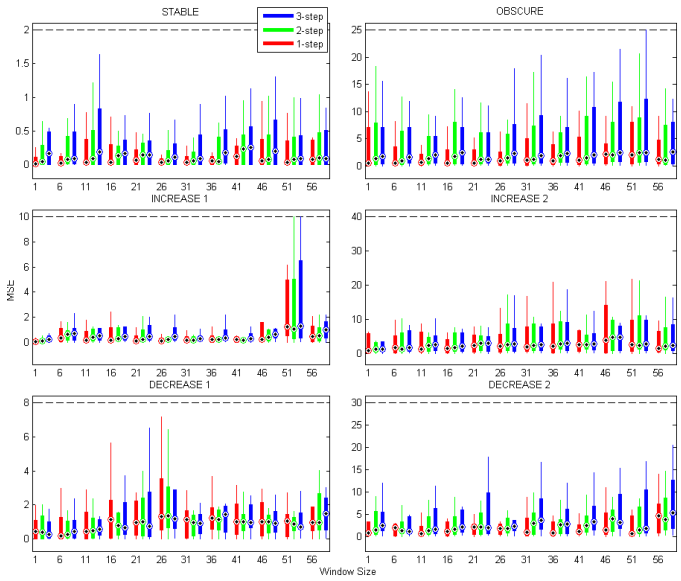


Fig. 4. Results of GPR Based Prediction (Different Window Sizes)

predicted. And also almost all the result ranges of window size 15 were the largest and window size 5 were the smallest. Thus, when the window size was small, reliable results could be obtained even when several data samples needed to be predicted. For OBSCURE case, about 80% of MSEs were below 5 for 1-step prediction. But when we tried to predict more steps ahead, the 75th percentile fluctuated around 10. For INCREASE 2, window size 5 and 10 had small MSEs when the prediction length was less than 5. Otherwise, the errors could be large and the predictions could be inaccurate. For DECREASE 2, all the three window sizes had the results where most errors fell below 10 for the prediction length no greater than 5. Then, as the prediction length increased, the MSEs would also increase.

Figures 4 presents the MSE values using sparse pseudo-input Gaussian process algorithm over different window sizes and the color boxes showed the results on different prediction lengths. The STABLE trend had the best performance for all the three steps predictions with all the results below 2. INCREASE 1 had a good performance with MSEs less than 3 except the window size of 61. DECREASE 1 also had a good performance with MSEs below 8. For OBSCURE trend, most MSEs of 1-step prediction were in the range below 5. For 2-step and 3-step predictions, most errors were below 10 when the window size was less than 41 for INCREASE 2. Most of the 1-step and 2-step prediction results were less than 5 for DECREASE 2. Also, the 3-step prediction seemed to be less accurate than 1-step and 2-step with most of MSEs less than 10.

Figure 5 shows the results of the MSE for different prediction lengths on window sizes of 5, 10 and 15. The STABLE state had the best performance. As the prediction length increased to 20, the MSEs were smaller than 3.5. The errors increased when we predicted more points ahead for trends INCREASE 1 and DECREASE 1. When the prediction length was smaller than 15, the MSEs were most likely to be less than 5 for both cases. For OBSCURE trend, the 75th

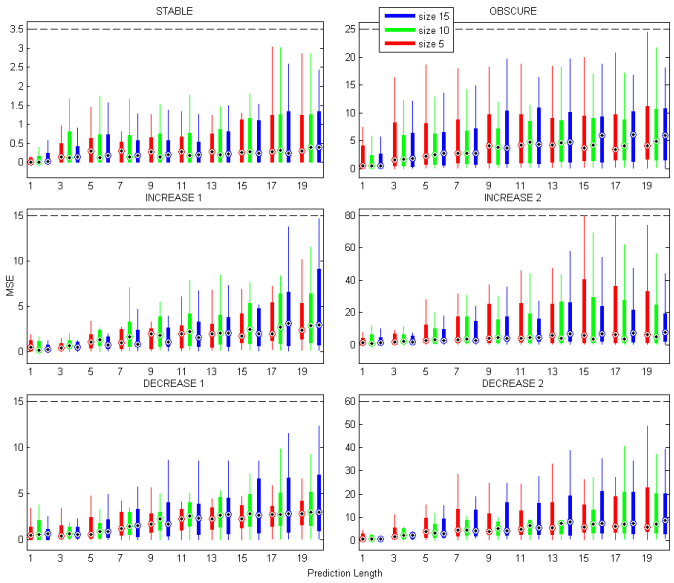


Fig. 5. Results of GPR Based Prediction (Different Prediction Lengths)

percentiles were mostly below 10 with a slowly increasing trend. For INCREASE 2, when predicting no more than 5 points ahead, the MSEs were less than 10. If we kept predicting, the errors would increase. For DECREASE 2, the predictions were accurate for all the three different window sizes to predict up to 5 data samples. Then, the error ranges increased slowly as prediction length increased. But most of the errors were below 20.

Figures 6 presents the MSE values using Kalman filter over different window sizes and different prediction lengths. The Kalman filter showed a good performance for STABLE, INCREASE 1 and DECREASE 1 trends. All the MSEs were less than 1.5, 1.5 and 3 for STABLE, INCREASE 1 and DECREASE 1 respectively. The 75th percentiles of MSEs were below 8 for OBSCURE, 6 for INCREASE 2 and 5

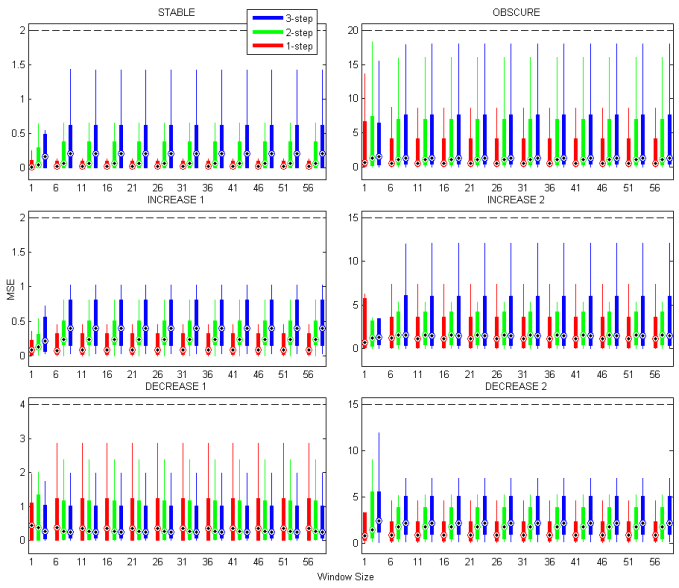


Fig. 6. Results of Kalman Filter Based Prediction (Different Window Sizes)

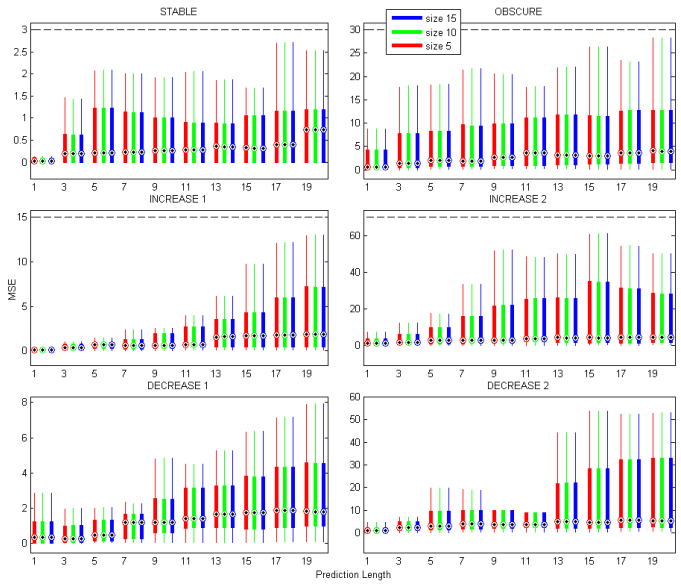


Fig. 7. Results of Kalman Filter Based Prediction (Different Prediction Lengths)

for DECREASE 2. Also the results did not change as the window size increased because the prediction only depended on the most current data and did not care about the previous information.

Figure 7 shows the distributions of the MSEs for different prediction lengths with window size of 5, 10 and 15. The STABLE trend had the best performance with all the errors less than 3. For INCREASE 1 and DECREASE 1, the ranges of MSEs increased as the prediction length increased. When the prediction length was less and equal than 15, the prediction errors were more likely to be less than 5 in INCREASE 1 trend. For DECREASE 1, most of the errors between the predictions and the actual values were below 5 when the prediction length was up to 20. The MSEs in OBSCURE were in the range less than 10 with a slowly increasing trend. Because the transition matrix A was set to be 1 so that $A^{(r-1)}$ was always 1 in this study, the MSEs were mainly obtained by the variations of the HR changes. For INCREASE 2, the predictions were very close to the expected values when the prediction length was less than 5. The error range increased when the prediction length reached 15. After this value, the error range decreased. Most of the MSEs were less than 10 when the prediction length was less and equal than 11 for DECREASE 2. Then the error range slowly increased as the prediction length increasing.

Since the KF-based prediction did not change for multi-step prediction and the mean imputation based prediction was using the average to replace all the missed data, we compared the performance of 1-step prediction for all three methods, shown in Figure 8. At STABLE trend, all the three methods had a good performance with MSEs less than 1. At INCREASE 1 and DECREASE 1 trends, both GPR and KF had the best performance with small errors. Mean imputation became very worse as the window size increasing. For INCREASE 2 and DECREASE 2, KF had the best performance and mean imputation was the worst. GPR also had a good performance. But when the window size was large, it could make more predictions far away from the actual data. For OBSCURE

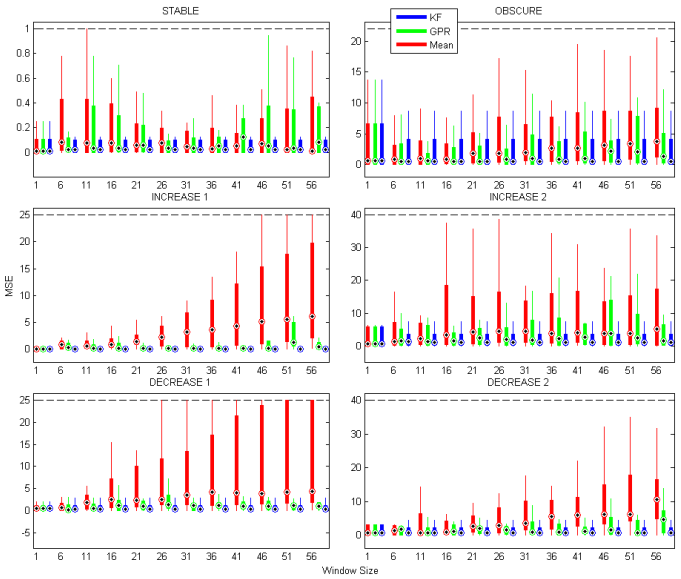


Fig. 8. 1-Step Prediction Comparison (KF, GPR, and Mean-based Approaches)

trend, in general, most MSEs were less than 10 for all the three methods and different window sizes. When the window size was less and equal than 26, GPR seemed to have the predictions more closer compared to other two methods. As window size increased, KF seemed to perform better than GPR.

IV. DISCUSSION

For all the three different prediction approaches, they had very good performance when the period data were in STABLE trend no matter how many past samples were considered.

For trend INCREASE 1 and DECREASE 1, both GPR and KF had good performance for the window size increasing from

1 to 56 with incremental of 5 on 1-step, 2-step and 3-step predictions. When GPR method dealing with the prediction problem, it first found the nonlinear line which fit largest data it could reached in the fixed window size. This nonlinear line showed the trend of the trained heart rates which was very helpful to predict data had the same trend information. Thus GPR showed small errors in predicting INCREASE 1 and DECREASE 1 trend data. Because the trend information held by the data in different window sizes was different, the predictions could be varied. Therefore, the ranges of results did not simple increasing as the window size increased. Since KF kept adjust its predictions by the error between the prediction and actual value in previous point, it was good to follow the INCREASE 1 and DECREASE 1 trends. So it could give predictions very close to the actual HRs. Also, KF only used the most recent measurement for prediction. This made the predictions constant although the window size increased. However, mean imputation only got small MSEs when the window size was small; otherwise, the errors increased a lot. The reason causing this was the characteristic of the data trend. Considered the INCREASE 1 trend for example since DECREASE 1 had the same reason. We averaged a fixed window size data before the starting point of predictions and used the mean value to replace the first prediction point. Since the trend was increasing, the mean value was absolutely less than what was expected. If more data were included for calculation, more smaller data were included. This would make the mean value even smaller than the actual value. Thus, mean imputation worked well only when the window size was small.

When the data had the trends of INCREASE 2 and DECREASE 2, it was hard for the prediction method to follow the immediate HR change. However, when the window size was small, the performance of all the three methods was similar with most of the MSEs below 5 which was a reasonable tolerant range. Increasing window size did not help to improve the performance. It even could have worse results.

For OBSCURE case, most of MSEs were below 10 for mean imputation, GPR and KF methods when the window size was small. Otherwise, the errors could increase.

It can happen that two or more data may be missing. For STABLE trend, the predictions, no matter how many samples we wanted to predict ahead, were very close to the actual HRs. For a fixed window size, mean imputation, GPR and KF had worse performance as more data were predicted. Because HRs were more relevant in a short time, the predictions would be less accurate for points several minutes away. For INCREASE 2 and DECREASE 2, all the three prediction methods only had small MSEs when the prediction length was small. For large prediction lengths, it was more likely to have less accurate predictions. For OBSCURE case, although the MSEs had a slight increasing, more than 75% errors were less than 10.

V. CONCLUSIONS

The missing data has been a long-standing issue in patient monitoring since it can introduce degraded signals and adversely affect the clinical decision making process. In this paper, we investigated three different approaches for missing data prediction in clinical settings: mean imputation, GPR and KF. Experimental results showed that KF and GPR had a good performance for 1-step prediction when the data were stable without frequently large changes; when dealing with data varying a lot frequently, the predicted values seemed to be less reliable. The larger window size did not necessarily mean the better performance. In most situations, the heart rate was more closely relevant to its most recent values.

REFERENCES

[1] S. Keehan, A. Sisko, C. Truffer, S. Smith, C. Cowan, J. Poisal, and M. K. Clemens. Health spending projections through 2017: The baby-boom generation is coming to medicare. *Health Affairs*, 27(2):w145–w155, 2008.

[2] OECD. *Health at a Glance 2013: OECD Indicators*. OECD Publishing, 2013.

[3] G. Tröster. The agenda of wearable healthcare. *IMIA Yearbook of Medical Informatics 2005: Ubiquitous Health Care Systems*, pages 125–138, 2005.

[4] G. D. Clifford, D. J. Scott, and M. Villarroel. User guide and documentation for the MIMIC II database. Feb 2012.

[5] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, Li-Wei Lehman, G. Moody, T. Heldt, Tin H. Kyaw, B. Moody, and R. G. Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): A public-access intensive care unit database. *Critical Care Medicine*, 39:952–960, May 2011.

[6] D. M. Korniewicz, T. Clark, and Y. David. A national online survey of the effectiveness of clinical alarms. *American Journal of Critical Care*, 17(1):36–41, Jan. 2008.

[7] P. C. A. Kam, A. C. Kam, and J. F. Thompson. Noise pollution in the anesthetic and intensive care environment. *Anaesthesia*, 49(11):982–986, Feb. 1994.

[8] M. Imhoff and S. Kuhls. Alarm algorithms in critical care monitoring. *Anesthesia & Analgesia*, 102(5):1525–1537, 2006.

[9] E. M. J. Koshi, A. Mäkivirta, T. Sukuvaara, and A. Kari. Frequency and reliability of alarms in the monitoring of cardiac postoperative patients. *J. Clinical Monitoring and Computing*, 7(2):129–133, 1990.

[10] S. T. Lawless. Crying wolf: false alarms in a pediatric intensive care unit. *Critical Care Medicine*, 22(6):981–985, Jun. 1994.

[11] A. Prez, R. J. Dennis, J. F. A. Gil, M. A. Rondn, and A Lpez. Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in colombia. *Statist. Med.*, 21:3885–3896, 2002.

[12] O.T. Abdala and M. Saeed. Estimation of missing values in clinical laboratory measurements of ICU patients using a weighted k-nearest neighbors algorithm. In *Computers in Cardiology, 2004*, pages 693–696, Sept 2004.

[13] G. B. Moody G. D. Clifford, W. J. Long and P. Szolovits. Robust parameter extraction for decision support using multimodal intensive care data. *Philosophical transactions of the Royal society, A*, 367(1887):411–429, Jan 2009.

[14] R.J.A. Little and D.B. Rubin. *Statistical Analysis With Missing Data*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 1987.

[15] CE Rasmussen and CKI Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, 1 2006.

[16] E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 1257–1264. MIT press, 2006.

[17] E. Snelson. *Flexible and efficient Gaussian process models for machine learning*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, London, 2007.

[18] R. G. Mark Q. Li and G. D. Clifford. Artificial arterial blood pressure artifact models and an evaluation of a robust blood pressure and heart rate estimator. *BioMedical Engineering OnLine*, 8:13–28, 2009.

[19] R. G. Mark Q. Li and G. D. Clifford. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter. *Physiol Meas.*, 29:15–32, 2008.