Gaussian process classification for prediction of in-hospital mortality among preterm infants

Olli-Pekka Rinta-Koski[a,*], Simo Särkkä[b], Jaakko Hollmén[a], Markus Leskinen[c], Sture Andersson[c]

[a] *Aalto University, Department of Computer Science, PO Box 15400, FI-00076 Aalto, Finland*

[b] *Aalto University, Department of Electrical Engineering and Automation, PO Box 12200, FI-00076 Aalto, Finland*

[c] *University of Helsinki, and Helsinki University Hospital, PO Box 140, FI-00029 HUS, Finland*

Abstract

We present a method for predicting preterm infant in-hospital mortality using Bayesian Gaussian process classification. We combined features extracted from sensor measurements, made during the first 72 hours of care for 598 Very Low Birth Weight infants of birth weight <1500 g, with standard clinical features calculated on arrival at the Neonatal Intensive Care Unit. Time periods of 12, 18, 24, 36, 48, and 72 hours were evaluated. We achieved a classification result with area under the receiver operating characteristic curve of 0.948, which is in excess of the results achieved by using the clinical standard SNAP-II and SNAPPE-II scores.

Keywords: time series prediction; Gaussian process classification; very low birth weight infants; neonatal intensive care

DOI: 10.1016/j.neucom.2017.12.064

[*] Corresponding author

Email addresses: `olli-pekka.rinta-koski@aalto.fi` (Olli-Pekka Rinta-Koski), `simo.sarkka@aalto.fi` (Simo Särkkä), `jaakko.hollmen@aalto.fi` (Jaakko Hollmén), `markus.leskinen@hus.fi` (Markus Leskinen), `sture.andersson@hus.fi` (Sture Andersson)

1. Introduction

This article is related to the use of data-driven methods in the context of digital healthcare and health informatics [1, 2]. In particular, our aim is to develop machine learning methodology for integration of heterogeneous data sources in order to more accurately predict the survival chances of preterm infants during treatment in the Neonatal Intensive Care Unit (NICU). First, we combine the conventional scoring system used in clinical practice with data-driven prediction from raw sensor data. Second, we study the prediction accuracy when the clinical scores are completely replaced with measurement data. The development of new methods for predicting neonatal in-hospital mortality is important, because while the global under-five mortality rate has dropped 53% since 1990, the proportion of neonatal deaths is projected to increase from 45% in 2015 to 52% by 2030 [3]. The incidence of certain complications (e.g. necrotizing enterocolitis) increases with the survival of preterm infants who previously would have died before the onset of these problems, emphasizing the need for developing new methods and strategies for neonatal intensive care [4]. Furthermore, data-only prediction is extremely important in clinical work, because the determination of the conventional scores is labor-intensive and requires that a specific set of diagnostic markers is available.

Routinely available markers of risk – sex, birth weight, and gestational age – fail to predict observed variation of mortality in NICUs [5]. This has prompted development of illness severity scores, such as SNAP-II and SNAPPE-II [6], which add laboratory results and physiological measurements of vital signs to perinatal risk factors in order to better predict morbidity and mortality. These risk scores were developed when patient records were mostly collected by hand, relying on simplified presentation of physiological data such as lowest temperature and mean blood pressure. Current patient information systems and patient monitors have made collection of detailed medical data much easier. We hypothesized that time series data of vital signs would help to identify patients at risk and, when combined with traditional risk scores, would result in increased predictive power.

The machine learning methodology that we use is based on the use of Gaussian process (GP) classification [7] with features extracted from raw cardiac, arterial and oximeter sensor measurements in addition to the clinical scores, gestational age at birth, and birth weight. Our motivation for studying GP classifiers in this context stems from two properties of GPs. First, they are genuine probabilistic models [7] and can provide information on how certain we are about the answer. This feature is inherent in GPs whereas, for example, for support vector machines (SVMs) [8] the uncertainty needs to be estimated with an additional model on the basic SVM [9]. Second, an even more important property is that GPs can flexibly be combined with first principles models [10, 11]. The resulting latent force models (LFMs) have a huge potential in medical applications especially due to their connection with time-series models used in sensor signal processing [12, 13, 14]. As shown in these papers, it is even possible to see that GPs models are solutions to certain stochastic partial differential equations, which not only allow for the combination with first-principles physical models, but also enable the use of Kalman filtering and other Bayesian filtering methods [15] for computationally efficient implementation of GP classifiers. GP classifiers have been previously used in health data analysis in (adult) Intensive Care Units (ICU) [16, 17, 18] and machine learning methods have been applied to NICU data [19, 20].

The contribution of this paper is that using cross-validation we show that augmenting the staff-determined SNAP-II and SNAPPE-II scores with sensor measurements improves prediction accuracy over standard clinical measures. We also show that a data-driven prediction from measurements alone can lead to better prediction accuracy than SNAP-II and SNAPPE-II. The proposed approach gives the area under the receiver operating characteristic curve (AUC) 0.946 for mortality prediction, which compares favourably with AUC 0.9151 reported for logistic regression by Saria et al. [20], and AUC 0.913 for CRIB-II and AUC 0.907 for SNAPPE-II reported by Reid et al. [21]. Although it has previously been shown [6] that in-hospital mortality of preterm infants is strongly correlated with birth weight and gestational age at birth, we show that the prediction result achieved by using these two variables alone (Table 2) can be improved by adding features extracted from measurement time series.

This article is an extended version of the conference article “Prediction of preterm infant mortality with Gaussian process classification” [22] presented at the 25th European Symposium on Articial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2017), in which we looked at data from the first postnatal 24 hours. Here the analysis has been extended to six different time periods ranging between the first 12 and 72 postnatal hours, three different kernels have been used with the GP classifier, and the classification performance has been compared to other classifiers.

2. Materials and methods

2.1. NICU database

The NICU at Helsinki University Hospital has been collecting patient data in a database since 1999. Data include measurements of clinical parameters such as oxygen saturation by pulse oximetry (SpO_2) and supplemental oxygen levels, observations made by staff, and clinical outcomes. Our study cohort includes 2059 Very Low Birth Weight (VLBW) infants (birth weight <1500 g) admitted between 1999–2013. Median gestational age at birth was 202 days (H28+6 weeks) and median birth weight was 1102 g.

The NICU database contains data recorded from equipment interfaces, as well as notations made by hand. Automatically gathered data consists of 111 different variables taken from monitor outputs of equipment used in the NICU. As the monitoring equipment and clinical guidelines have varied during the 15 year period under which the data has been stored, not all data is available for all 2059 patients.

2.2. Preprocessing and feature extraction

For the experiment, we decided to study the first 72 hours from delivery to see whether the time series data gathered during that period has predictive power. Most in-hospital deaths occur within the first week; median in this dataset is 5 days. There are 598 patients in the dataset for whom there is complete data from the first 72 hours of their NICU stay for each of these seven variables: gestational age at birth, birth weight, systolic, mean, and diastolic arterial blood pressure, heart rate measured by electrocardiography (ECG), and SpO_2. If for some sensor signal there were only a few measurements available, the patient data was considered incomplete. Patients that died before the end of 72 hour period were excluded as well. The in-hospital mortality rate of this subset is 9% (53 patients), which is also the mortality rate in the full cohort. In addition to the full 72 hour period, we also looked at the first 12, 18, 24, 36, and 48 hour periods.

The data was preprocessed by removing out-of-range values caused by, for example, misplaced or removed sensors and monitoring equipment drifting out of calibration from the time series.

For feature extraction, mean and standard deviation were calculated from each of the following time series for each patient: systolic, mean, and arterial blood pressure, ECG heart rate, and SpO_2. SNAP-II score, SNAPPE-II score, gestational age at birth, and birth weight were directly used as features. We chose not to use any more complicated features such as signal derivatives, because the signals streams were very sparse and noisy, and reliably estimating the signal derivatives would have required us to use Kalman filter type of methods [15], which we wanted to avoid at this stage in order to keep the preprocessing simple and robust.

2.3. Gaussian process classifier

We used a GP [7] classifier with a probit measurement model:

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')), \quad p(y_i \mid f(\mathbf{x}_i)) = \int_{-\infty}^{y_i f(\mathbf{x}_i)} \mathrm{N}(z \mid 0, 1)\, dz, \tag{1}$$

where the classes are labeled as $y_i \in \{-1, 1\}$. This choice of the measurement model is standard in GP literature [7] and is supported by most GP software packages such as the GPstuff Toolbox [23].

The kernel was a sum of squared exponential (or radial basis function) kernel, linear kernel, and constant kernel:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_{\mathrm{se}}^2 \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}')^\mathsf{T}\,\Lambda^{-1}\,(\mathbf{x}-\mathbf{x}')\right) + \mathbf{x}^\mathsf{T}\,\Sigma\,\mathbf{x}' + \sigma^2, \tag{2}$$

where $\Lambda = \mathrm{diag}(l_1^2, \ldots, l_d^2)$ and $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. The rationale behind this kernel choice is that the constant and the linear parts of the kernel aim at capturing the bias and the linear trend in the problem, respectively. In order to capture the non-linear effects, we add the squared exponential kernel with the automatic relevance determination prior, which is a commonly used general covariance function in Gaussian process regression [7]. This kind of 3-part covariance functions have also been recently used in medical applications [24, 25].

For comparison purposes, we also used the Matérn kernel with $\nu = 3/2$ (M32) and $\nu = 5/2$ (M52) [7, 23, 24, 25] replacing the squared exponential kernel in the 3-part kernel:

$$k_{\nu=3/2}(\mathbf{x}, \mathbf{x}') = \sigma_{\mathrm{m}}^2(1 + \sqrt{3}\|\mathbf{x}-\mathbf{x}'\|)\exp(-\sqrt{3}\|\mathbf{x}-\mathbf{x}'\|) \tag{3}$$

$$k_{\nu=5/2}(\mathbf{x}, \mathbf{x}') = \sigma_{\mathrm{m}}^2(1 + \sqrt{5}\|\mathbf{x}-\mathbf{x}'\| + \frac{5\|\mathbf{x}-\mathbf{x}'\|^2}{3})\exp(-\sqrt{5}\|\mathbf{x}-\mathbf{x}'\|) \tag{4}$$

For training the classifier we used the GPstuff Toolbox [23] with Laplace approximation on the latent variables and circular composite design (CCD) integration over the hyperparameters. The CCD method has advantages over, for example, marginal likelihood maximization due it better handling of uncertainty. In particular, the method approximates the integration over the hyperparameters instead of using a plug-in point-estimate, which ensures that the uncertainly is computed in a proper Bayesian way.

In order to evaluate the performance of the classifiers we used stratified 8-fold cross-validation (CV) which takes the class priors into account when forming the partitions. Cross-validation was used to estimate the classification accuracy, precision, specificity, and sensitivity as well as receiver operating characteristic (ROC) curve [26] and the area under the ROC curve (AUC) [27]. In order to reduce the variance of CV, we repeated each CV run 8 times and averaged the results.

2.4. Comparison with other classifiers

We used the following classifiers in comparison with the GP classifier:

- *SNAP-II/SNAPPE-II thresholding.* Thresholding using only the SNAP-II or SNAPPE-II scores (one at a time) was used to classify the patients. The class boundary was set using one of two rules. In the first case, the maximum accuracy achieved with the training set was used to set the class boundary. In the second case, the maximum value of the Youden index [28] was used. This gave us four different rule-score combinations.
- *Support vector machine classifier.* A linear SVM classifier [8] was used as the classifier and the posterior probability estimates were obtained with Platt scaling [9]. The ROC curve was calculated by sweeping the class boundary from 0 to 1. The prediction was given by setting the class boundary to 0.5.
- *Linear probit model.* A linear model with a probit link function was implemented by using a constant plus a linear kernel in a GP classification model. The model was trained using the GPstuff Toolbox. The integration over the hyperparameters was performed using the CCD method [29].
- *Random classifier.* This classifier assigns the class at random weighted by class prior probabilities of the training set.
- *Majority classifier.* This classifier simply assumes that all patients belong to the larger (survivor) class.

3. Results

3.1. Classification with SNAP-II and SNAPPE-II scores

First, we tested the performance of the classifiers using only SNAP-II and SNAPPE-II scores with gestational age at birth and birth weight. Although this information is equal to what the scores are traditionally computed from, as can be seen in Tables 1 and 2, the GP classifier is able to achieve a better

AUC (0.933) than the clinical standard SNAP-II (AUC 0.860) and SNAPPE-II (AUC 0.878) scores, with all variants (sum, M32, M52) giving practically the same result and linear probit classifier at a just slightly lower AUC (0.921).

Table 1: Reference results. SNAP/SNAPPE = SNAP-II/SNAPPE-II with optimal (cross-validated) thresholding (A = maximal accuracy, Y = Youden index), Majority = trivial classifier that assumes all patients survive, Random = class picked at random weighted by training set class priors. Acc = accuracy, PPV = positive predictive value, Sens = sensitivity, Spec = specificity, AUC = area under the receiver operating characteristic curve. Values in parentheses indicate the associated standard error. Results in all tables in descending order by AUC.

	Acc	PPV	Sens	Spec	AUC
SNAPPE A	0.914 (0.00)	0.898 (0.05)	0.056 (0.02)	0.998 (0.00)	0.875 (0.00)
SNAPPE Y	0.737 (0.01)	0.248 (0.01)	0.923 (0.02)	0.719 (0.01)	0.875 (0.00)
SNAP A	0.909 (0.00)	0.677 (0.08)	0.062 (0.02)	0.993 (0.00)	0.859 (0.00)
SNAP Y	0.713 (0.01)	0.227 (0.01)	0.895 (0.02)	0.695 (0.01)	0.859 (0.00)
Random	0.839 (0.01)	0.091 (0.02)	0.091 (0.02)	0.913 (0.01)	0.500 (0.00)
Majority	0.910 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.500 (0.00)

Table 2: Results using only SNAP-II, SNAPPE-II, gestational age at birth, and birth weight.

	Acc	PPV	Sens	Spec	AUC
GP M32	0.918 (0.00)	0.611 (0.04)	0.360 (0.03)	0.974 (0.00)	0.933 (0.00)
GP	0.919 (0.00)	0.618 (0.05)	0.351 (0.03)	0.975 (0.00)	0.933 (0.00)
GP M52	0.919 (0.00)	0.615 (0.04)	0.358 (0.03)	0.974 (0.00)	0.933 (0.00)
Linear	0.914 (0.00)	0.579 (0.05)	0.260 (0.03)	0.978 (0.00)	0.921 (0.00)
SVM	0.907 (0.00)	0.857 (0.06)	0.020 (0.01)	0.994 (0.00)	0.644 (0.01)

Next, we used all available signals with the GP classifier in order to get an upper bound on the achievable performance. All the available features were used as classifier inputs, in other words, SNAP-II, SNAPPE-II, gestational age at birth, birth weight, and the mean and standard deviation of each of the following: systolic, mean, and diastolic arterial blood pressure, ECG heart rate, and SpO_2.

Table 3 (all available features) shows GP prediction results using all available features with three different kernels (sum, M32, and M52). Kernel choice had a negligible effect. The highest AUC (0.948) was achieved with 48h data and the sum kernel. All AUC values from predictions with all three kernels for time periods between 36h and 72h were within 0.007. Shortening the range of time series data has a slight negative effect on the AUC values, with 12h data and the sum kernel yielding AUC 0.924. However, as the range decreases, there is a drop in both positive predictive value (PPV), from 0.708 to 0.598, and sensitivity, from 0.463 to 0.283. SVM and the linear probit model give similar results to GP (Tables 4 and 5) in many of the cases, but with shorter ranges the GP models give slightly better results.

Table 3: GP prediction results using all available features and three different kernels (sum, M32, M52).

	Acc	PPV	Sens	Spec	AUC
48h GP	0.930 (0.00)	0.660 (0.03)	0.463 (0.02)	0.975 (0.00)	0.948 (0.00)
48h GPm32	0.928 (0.00)	0.649 (0.03)	0.445 (0.02)	0.975 (0.00)	0.947 (0.00)
48h GPm52	0.928 (0.00)	0.657 (0.03)	0.442 (0.02)	0.976 (0.00)	0.946 (0.00)
36h GPm32	0.925 (0.00)	0.667 (0.03)	0.391 (0.02)	0.977 (0.00)	0.945 (0.00)
72h GP	0.932 (0.00)	0.708 (0.03)	0.449 (0.02)	0.980 (0.00)	0.942 (0.01)
72h GPm52	0.933 (0.00)	0.717 (0.03)	0.453 (0.02)	0.980 (0.00)	0.942 (0.01)
36h GPm52	0.925 (0.00)	0.669 (0.03)	0.390 (0.02)	0.977 (0.00)	0.942 (0.01)
72h GPm32	0.934 (0.00)	0.727 (0.03)	0.452 (0.02)	0.981 (0.00)	0.942 (0.01)
36h GP	0.924 (0.00)	0.670 (0.03)	0.389 (0.02)	0.977 (0.00)	0.941 (0.01)
24h GPm32	0.918 (0.00)	0.591 (0.04)	0.332 (0.03)	0.976 (0.00)	0.931 (0.01)
24h GPm52	0.919 (0.00)	0.596 (0.04)	0.331 (0.03)	0.977 (0.00)	0.930 (0.01)
18h GPm32	0.920 (0.00)	0.671 (0.03)	0.312 (0.02)	0.981 (0.00)	0.929 (0.01)
24h GP	0.919 (0.00)	0.624 (0.03)	0.335 (0.02)	0.976 (0.00)	0.929 (0.00)
18h GP	0.920 (0.00)	0.655 (0.04)	0.328 (0.02)	0.979 (0.00)	0.928 (0.01)
18h GPm52	0.922 (0.00)	0.682 (0.03)	0.335 (0.02)	0.980 (0.00)	0.928 (0.01)
12h GP	0.915 (0.00)	0.598 (0.04)	0.283 (0.02)	0.977 (0.00)	0.924 (0.01)
12h GPm32	0.913 (0.00)	0.590 (0.04)	0.295 (0.02)	0.974 (0.00)	0.923 (0.01)
12h GPm52	0.913 (0.00)	0.581 (0.03)	0.297 (0.02)	0.974 (0.00)	0.921 (0.01)

Finally, Table 6 shows the results for all non-reference classifiers using all available features. GP kernel choice had negligible effect. The linear probit model performs worse than GP with 12h and 18h data. With longer time series, GP and the linear probit model have roughly equal performance. Even the lowest AUC (0.901), given by the SVM classifier with 12h data, is better than SNAP-II/SNAPPE-II thresholding (AUC 0.859. . . 0.875).

Table 4: SVM prediction results using all available features.

	Acc	PPV	Sens	Spec	AUC
36h	0.931 (0.00)	0.708 (0.03)	0.443 (0.02)	0.979 (0.00)	0.947 (0.00)
48h	0.931 (0.00)	0.725 (0.03)	0.433 (0.02)	0.980 (0.00)	0.943 (0.00)
24h	0.923 (0.00)	0.665 (0.03)	0.340 (0.02)	0.981 (0.00)	0.930 (0.01)
72h	0.931 (0.00)	0.746 (0.03)	0.384 (0.02)	0.984 (0.00)	0.924 (0.01)
18h	0.918 (0.00)	0.692 (0.04)	0.259 (0.02)	0.984 (0.00)	0.922 (0.00)
12h	0.910 (0.00)	0.564 (0.04)	0.182 (0.02)	0.982 (0.00)	0.901 (0.01)

Table 5: Linear model prediction results using all available features.

	Acc	PPV	Sens	Spec	AUC
48h	0.926 (0.00)	0.645 (0.02)	0.463 (0.02)	0.971 (0.00)	0.949 (0.00)
36h	0.924 (0.00)	0.652 (0.03)	0.402 (0.02)	0.975 (0.00)	0.949 (0.00)
72h	0.934 (0.00)	0.720 (0.03)	0.475 (0.02)	0.979 (0.00)	0.944 (0.00)
24h	0.919 (0.00)	0.610 (0.04)	0.348 (0.02)	0.976 (0.00)	0.931 (0.01)
18h	0.922 (0.00)	0.675 (0.04)	0.326 (0.02)	0.981 (0.00)	0.927 (0.01)
12h	0.913 (0.00)	0.583 (0.04)	0.279 (0.02)	0.976 (0.00)	0.916 (0.01)

3.2. Classification with reduced feature sets

To find out how the classifiers perform with reduced feature sets, we tested the classifiers without SNAP-II and SNAPPE-II scores (Table 7) and finally with sensor signals only (dropping also gestational age at birth and birth weight, Table 8).

Without SNAP-II/SNAPPE-II, the linear probit model and GP perform equally well with time periods of at least 36h (AUC 0.943. . . 0.946). All classifiers outperform the reference results (Table 1), with the exception of SVM with 12h data (AUC 0.874) which achieves a result comparable with SNAPPE-II thresholding.

Table 8 shows prediction results using only time series data. The best classifier is GP (all kernels) with 48h data (AUC 0.925. . . 0.926) but linear classifiers with 48h and 72h data as well as GP with 72h data perform almost equally well. Whereas the GP kernel choice is again practically immaterial, both AUC and sensitivity increase as the time series grows longer, AUC from 0.787 (12h data, M32 kernel) to 0.926 (48h data, sum kernel). Interestingly, 72h data gives slightly lower AUC values than 48h data (AUC 0.915. . . 0.919 vs. 0.925. . . 0.926),

Table 6: Prediction results using all available features. Comparison of GP, SVM, and linear probit model.

	Acc	PPV	Sens	Spec	AUC
48h Linear	0.926 (0.00)	0.645 (0.02)	0.463 (0.02)	0.971 (0.00)	0.949 (0.00)
36h Linear	0.924 (0.00)	0.652 (0.03)	0.402 (0.02)	0.975 (0.00)	0.949 (0.00)
48h GP	0.930 (0.00)	0.660 (0.03)	0.463 (0.02)	0.975 (0.00)	0.948 (0.00)
36h SVM	0.931 (0.00)	0.708 (0.03)	0.443 (0.02)	0.979 (0.00)	0.947 (0.00)
48h GPm32	0.928 (0.00)	0.649 (0.03)	0.445 (0.02)	0.975 (0.00)	0.947 (0.00)
48h GPm52	0.928 (0.00)	0.657 (0.03)	0.442 (0.02)	0.976 (0.00)	0.946 (0.00)
36h GPm32	0.925 (0.00)	0.667 (0.03)	0.391 (0.02)	0.977 (0.00)	0.945 (0.00)
72h Linear	0.934 (0.00)	0.720 (0.03)	0.475 (0.02)	0.979 (0.00)	0.944 (0.00)
48h SVM	0.931 (0.00)	0.725 (0.03)	0.433 (0.02)	0.980 (0.00)	0.943 (0.00)
72h GP	0.932 (0.00)	0.708 (0.03)	0.449 (0.02)	0.980 (0.00)	0.942 (0.01)
72h GPm52	0.933 (0.00)	0.717 (0.03)	0.453 (0.02)	0.980 (0.00)	0.942 (0.01)
36h GPm52	0.925 (0.00)	0.669 (0.03)	0.390 (0.02)	0.977 (0.00)	0.942 (0.01)
72h GPm32	0.934 (0.00)	0.727 (0.03)	0.452 (0.02)	0.981 (0.00)	0.942 (0.01)
36h GP	0.924 (0.00)	0.670 (0.03)	0.389 (0.02)	0.977 (0.00)	0.941 (0.01)
24h GPm32	0.918 (0.00)	0.591 (0.04)	0.332 (0.03)	0.976 (0.00)	0.931 (0.01)
24h Linear	0.919 (0.00)	0.610 (0.04)	0.348 (0.02)	0.976 (0.00)	0.931 (0.01)
24h SVM	0.923 (0.00)	0.665 (0.03)	0.340 (0.02)	0.981 (0.00)	0.930 (0.01)
24h GPm52	0.919 (0.00)	0.596 (0.04)	0.331 (0.03)	0.977 (0.00)	0.930 (0.01)
18h GPm32	0.920 (0.00)	0.671 (0.03)	0.312 (0.02)	0.981 (0.00)	0.929 (0.01)
24h GP	0.919 (0.00)	0.624 (0.03)	0.335 (0.02)	0.976 (0.00)	0.929 (0.00)
18h GP	0.920 (0.00)	0.655 (0.04)	0.328 (0.02)	0.979 (0.00)	0.928 (0.01)
18h GPm52	0.922 (0.00)	0.682 (0.03)	0.335 (0.02)	0.980 (0.00)	0.928 (0.01)
18h Linear	0.922 (0.00)	0.675 (0.04)	0.326 (0.02)	0.981 (0.00)	0.927 (0.01)
72h SVM	0.931 (0.00)	0.746 (0.03)	0.384 (0.02)	0.984 (0.00)	0.924 (0.01)
12h GP	0.915 (0.00)	0.598 (0.04)	0.283 (0.02)	0.977 (0.00)	0.924 (0.01)
12h GPm32	0.913 (0.00)	0.590 (0.04)	0.295 (0.02)	0.974 (0.00)	0.923 (0.01)
18h SVM	0.918 (0.00)	0.692 (0.04)	0.259 (0.02)	0.984 (0.00)	0.922 (0.00)
12h GPm52	0.913 (0.00)	0.581 (0.03)	0.297 (0.02)	0.974 (0.00)	0.921 (0.01)
12h Linear	0.913 (0.00)	0.583 (0.04)	0.279 (0.02)	0.976 (0.00)	0.916 (0.01)
12h SVM	0.910 (0.00)	0.564 (0.04)	0.182 (0.02)	0.982 (0.00)	0.901 (0.01)

but with better PPV and sensitivity (PPV 0.804. . . 0.813 vs. 0.639. . . 0.645, sensitivity 0.347. . . 0.361 vs. 0.320. . . 0.335). Time periods shorter than 36h do not give better than reference results with any of the classifiers.

The best SVM result (AUC 0.899, 48h data) equals the performance of GP and linear classifiers with 36h data, but loses to both with time periods of 48h and 72h. SVM performance degrades markedly with 24h and shorter data, not beating even the reference (SNAP-II/SNAPPE-II) classifiers.

Table 7: Prediction results using all available features except SNAP-II and SNAPPE-II.

	Acc	PPV	Sens	Spec	AUC
48h GPm32	0.928 (0.00)	0.680 (0.03)	0.442 (0.02)	0.976 (0.00)	0.947 (0.00)
48h Linear	0.926 (0.00)	0.655 (0.03)	0.464 (0.02)	0.971 (0.00)	0.947 (0.00)
48h GP	0.929 (0.00)	0.678 (0.03)	0.445 (0.02)	0.976 (0.00)	0.946 (0.00)
48h GPm52	0.926 (0.00)	0.668 (0.03)	0.433 (0.02)	0.975 (0.00)	0.946 (0.00)
72h GPm32	0.934 (0.00)	0.735 (0.02)	0.445 (0.02)	0.981 (0.00)	0.946 (0.00)
36h GP	0.922 (0.00)	0.634 (0.03)	0.400 (0.02)	0.973 (0.00)	0.945 (0.00)
36h GPm52	0.922 (0.00)	0.617 (0.03)	0.403 (0.02)	0.973 (0.00)	0.945 (0.00)
36h GPm32	0.923 (0.00)	0.627 (0.03)	0.412 (0.02)	0.973 (0.00)	0.945 (0.00)
72h Linear	0.934 (0.00)	0.705 (0.02)	0.503 (0.02)	0.976 (0.00)	0.945 (0.00)
36h Linear	0.927 (0.00)	0.655 (0.03)	0.457 (0.02)	0.973 (0.00)	0.945 (0.00)
72h GPm52	0.933 (0.00)	0.723 (0.02)	0.448 (0.02)	0.980 (0.00)	0.944 (0.01)
72h GP	0.932 (0.00)	0.713 (0.03)	0.439 (0.02)	0.980 (0.00)	0.943 (0.01)
48h SVM	0.927 (0.00)	0.691 (0.03)	0.413 (0.02)	0.978 (0.00)	0.941 (0.00)
36h SVM	0.930 (0.00)	0.699 (0.03)	0.413 (0.02)	0.981 (0.00)	0.941 (0.00)
24h Linear	0.924 (0.00)	0.639 (0.03)	0.396 (0.03)	0.976 (0.00)	0.931 (0.00)
24h GP	0.921 (0.00)	0.608 (0.03)	0.357 (0.03)	0.977 (0.00)	0.930 (0.00)
24h GPm52	0.919 (0.00)	0.590 (0.03)	0.349 (0.03)	0.975 (0.00)	0.930 (0.00)
24h GPm32	0.918 (0.00)	0.570 (0.04)	0.343 (0.03)	0.975 (0.00)	0.930 (0.00)
18h GPm52	0.921 (0.00)	0.674 (0.03)	0.328 (0.02)	0.980 (0.00)	0.925 (0.01)
18h GPm32	0.921 (0.00)	0.670 (0.03)	0.333 (0.02)	0.980 (0.00)	0.925 (0.01)
18h GP	0.923 (0.00)	0.678 (0.03)	0.345 (0.02)	0.981 (0.00)	0.924 (0.01)
72h SVM	0.932 (0.00)	0.742 (0.03)	0.405 (0.02)	0.984 (0.00)	0.923 (0.01)
18h Linear	0.921 (0.00)	0.659 (0.03)	0.337 (0.02)	0.979 (0.00)	0.920 (0.01)
24h SVM	0.924 (0.00)	0.733 (0.03)	0.306 (0.02)	0.985 (0.00)	0.919 (0.01)
12h GPm32	0.914 (0.00)	0.586 (0.04)	0.258 (0.02)	0.979 (0.00)	0.914 (0.01)
12h GPm52	0.913 (0.00)	0.591 (0.04)	0.253 (0.02)	0.979 (0.00)	0.913 (0.01)
12h GP	0.912 (0.00)	0.562 (0.04)	0.253 (0.02)	0.977 (0.00)	0.912 (0.01)
18h SVM	0.921 (0.00)	0.738 (0.03)	0.259 (0.02)	0.987 (0.00)	0.907 (0.01)
12h Linear	0.915 (0.00)	0.594 (0.04)	0.256 (0.02)	0.981 (0.00)	0.900 (0.01)
12h SVM	0.909 (0.00)	0.622 (0.05)	0.084 (0.01)	0.990 (0.00)	0.874 (0.01)

3.3. The effect of varying input combinations and time series lengths

Figure 1 shows the ROC curves for classifier results using all features, without SNAP-II/SNAPPE-II, and time series data only for the two extremes (12h, 72h) of time series lengths. The classification performance is improved with longer time series in all three cases.

Table 8: Prediction results using only time series data.

	Acc	PPV	Sens	Spec	AUC
48h GP	0.923 (0.00)	0.640 (0.03)	0.335 (0.02)	0.981 (0.00)	0.926 (0.00)
48h GPm52	0.923 (0.00)	0.639 (0.04)	0.322 (0.02)	0.981 (0.00)	0.925 (0.00)
48h GPm32	0.923 (0.00)	0.645 (0.04)	0.320 (0.02)	0.982 (0.00)	0.925 (0.00)
72h GPm32	0.932 (0.00)	0.813 (0.03)	0.347 (0.02)	0.989 (0.00)	0.919 (0.00)
72h GPm52	0.933 (0.00)	0.820 (0.03)	0.361 (0.02)	0.989 (0.00)	0.917 (0.00)
48h Linear	0.922 (0.00)	0.674 (0.04)	0.315 (0.02)	0.982 (0.00)	0.917 (0.00)
72h GP	0.932 (0.00)	0.804 (0.03)	0.352 (0.02)	0.989 (0.00)	0.915 (0.00)
72h Linear	0.933 (0.00)	0.798 (0.03)	0.360 (0.02)	0.988 (0.00)	0.913 (0.00)
36h GPm52	0.922 (0.00)	0.715 (0.04)	0.261 (0.02)	0.986 (0.00)	0.902 (0.01)
36h GPm32	0.922 (0.00)	0.716 (0.04)	0.259 (0.02)	0.987 (0.00)	0.902 (0.01)
36h GP	0.923 (0.00)	0.734 (0.04)	0.266 (0.02)	0.988 (0.00)	0.900 (0.01)
48h SVM	0.922 (0.00)	0.747 (0.04)	0.261 (0.02)	0.986 (0.00)	0.899 (0.01)
36h Linear	0.923 (0.00)	0.740 (0.04)	0.272 (0.02)	0.987 (0.00)	0.898 (0.01)
72h SVM	0.932 (0.00)	0.863 (0.03)	0.287 (0.02)	0.994 (0.00)	0.892 (0.01)
36h SVM	0.925 (0.00)	0.874 (0.03)	0.213 (0.02)	0.995 (0.00)	0.881 (0.01)
24h GPm32	0.918 (0.00)	0.701 (0.05)	0.194 (0.02)	0.990 (0.00)	0.868 (0.01)
24h GPm52	0.917 (0.00)	0.690 (0.05)	0.194 (0.02)	0.989 (0.00)	0.864 (0.01)
24h GP	0.919 (0.00)	0.703 (0.04)	0.208 (0.02)	0.989 (0.00)	0.857 (0.01)
24h Linear	0.917 (0.00)	0.665 (0.05)	0.186 (0.02)	0.989 (0.00)	0.857 (0.01)
18h GPm52	0.921 (0.00)	0.788 (0.04)	0.191 (0.02)	0.994 (0.00)	0.846 (0.01)
18h GPm32	0.920 (0.00)	0.796 (0.04)	0.181 (0.02)	0.994 (0.00)	0.845 (0.01)
18h GP	0.921 (0.00)	0.789 (0.04)	0.186 (0.02)	0.994 (0.00)	0.844 (0.01)
18h Linear	0.914 (0.00)	0.714 (0.04)	0.145 (0.02)	0.991 (0.00)	0.831 (0.01)
12h GP	0.914 (0.00)	0.737 (0.05)	0.101 (0.01)	0.994 (0.00)	0.799 (0.01)
24h SVM	0.912 (0.00)	0.798 (0.04)	0.089 (0.01)	0.994 (0.00)	0.793 (0.01)
12h GPm52	0.914 (0.00)	0.734 (0.05)	0.102 (0.01)	0.993 (0.00)	0.791 (0.01)
12h GPm32	0.913 (0.00)	0.727 (0.05)	0.101 (0.01)	0.993 (0.00)	0.787 (0.01)
12h Linear	0.910 (0.00)	0.701 (0.05)	0.076 (0.01)	0.992 (0.00)	0.779 (0.01)
12h SVM	0.910 (0.00)	0.969 (0.02)	0.000 (0.00)	0.999 (0.00)	0.669 (0.02)
18h SVM	0.909 (0.00)	0.935 (0.03)	0.012 (0.01)	0.999 (0.00)	0.597 (0.02)

Figure 2 shows the effect of varying the length of the time series. Increasing time series length improves the prediction result up to 48h for GP and the linear model. There is no marked difference between the 48h and 72h predictions. SVM performance peaks at 32h.

Using only time series data AUC is initially low, surpassing that of the SNAP-II/SNAPPE-II combination with 36h and longer time series. The best AUC was achieved with 48h data. Using the full 72h time series results in

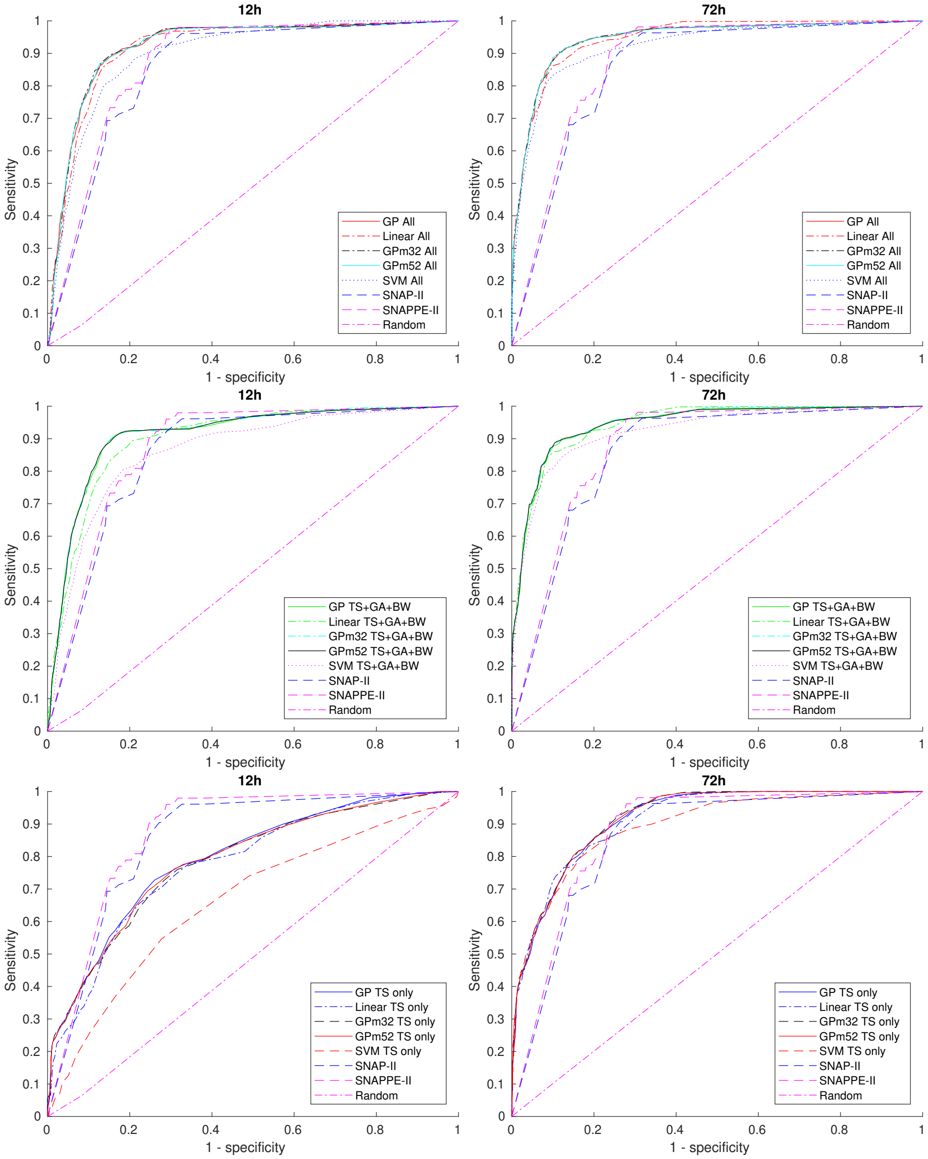

Figure 1: ROC curves for 12h and 72h classifiers. The ROCs of SNAP-II, SNAPPE-II, and the random classifier are also shown. Top row: all available features. Middle row: time series data + GA + BW. Bottom row: time series data only.

slightly lower AUC scores. The addition of GA and BW improves the result with short time series (12h and 18h), but has little effect with 24h and longer

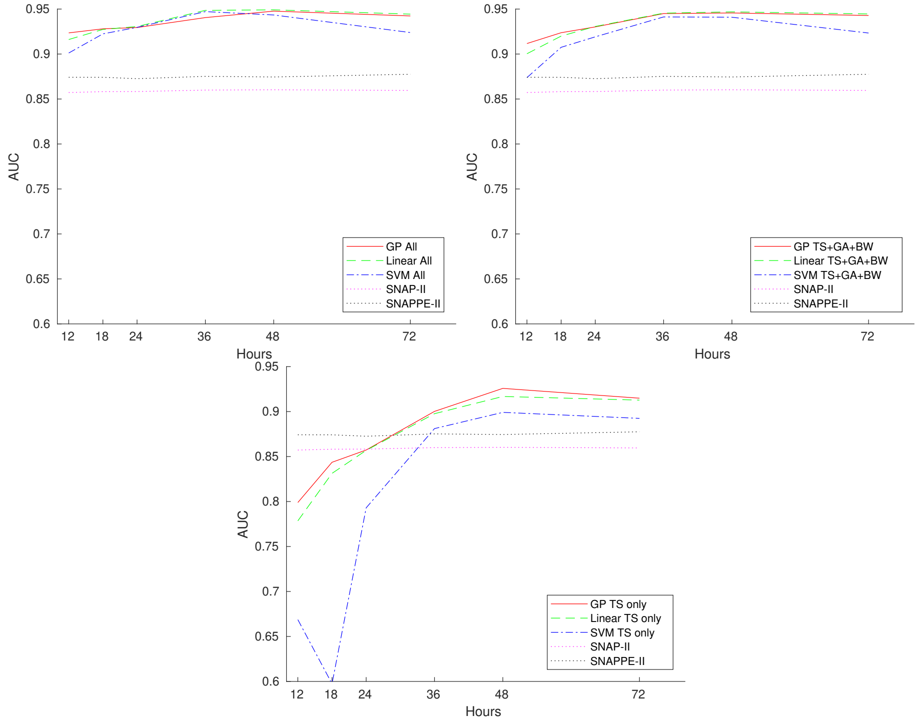

Figure 2: AUC values for different time series lengths. Top left: all variables. Top right: time series data with GA and BW. Bottom: time series data only.

time series.

4. Discussion

In another study [30], birth weight alone was found to have AUC 0.74 and gestational age alone had AUC 0.71. Both were inferior to the Clinical Risk Index for Babies (CRIB) [31], which had AUC 0.82. CRIB-II [32] was also found to have inferior predictive power at AUC 0.69. A comparison study of CRIB-II and SNAPPE-II [21] found the two scores performing equally well, with CRIB-II AUC 0.913 (SE 0.014) and SNAPPE-II AUC 0.907 (SE 0.012), while another comparison study [33] found CRIB (AUC 0.90) and CRIB-II (AUC 0.91) superior to SNAPPE-II (AUC 0.84, which is close to our SNAPPE-II thresholding result).

While our best prediction results were achieved using all available variables, adding SNAP-II and SNAPPE-II to time series data with gestational age and birth weight did not markedly improve the results. This is not surprising given that both scores are influenced by GA and BW to a great extent.

Both GP classification and the linear probit model gave practically identical results with time series of 36h and longer. With time series data only, GP and the linear probit model lose to SNAP-II/SNAPPE-II thresholding with 12h and 18h data, achieve roughly equal performance with 24h data, and beat them with 36h and longer time series, as does SVM. However, SVM performs significantly worse than SNAP-II/SNAPPE-II with 12h to 24h data.

In Figure 2 there is a slight drop in performance when using 72h data instead of 48h data. Although this may look surprising, it could be explained by the fact that the feature computations did not explicitly take the length of the time interval into account. It is thus possible that by, for example, liming the feature computations to the end of the time series or by taking the length of the time interval otherwise into account could improve the predictions.

The highest sensitivity achieved was 0.475 for the linear probit model using all variables with 72h data. GP sensitivies varied from 0.283 to 0.453 using all variables, dropping down markedly (0.101. . . 0.361) with only time series data. It is worth noting that low sensitivities of predictions do not necessarily mean that the clinical value of the predictions is low. From the clinical viewpoint, specificity is more important than sensitivity when predicting mortality. If the clinicians suspect that there is a high risk that the preterm infant will die, this can affect decisions to perform risky operations or start resource-intensive treatments. These kinds of decisions require careful consideration of the clinical situation and never rely on a single factor, such as predictive models. The goal is to have as high specificity as possible to avoid withholding treatment.

The prediction of in-hospital death in itself is not something that would be a major factor in how to treat the patient, but it can be useful in deciding whether to use some heavy means of care such as complex operations which themselves can be a risk to the patient. For that reason we have chosen to use data from the early phase of the NICU stay. In the early stages the medical personnel have not yet been able to form a complete view of the patient's state.

5. Conclusions

Time series data from the initial hours of a preterm infant's intensive care unit stay can be used to improve the accuracy of existing methods for predicting in-hospital death. A Bayesian Gaussian process classifier can be used to create a predictive model. Combining features extracted from time series data with clinical scores calculated on arrival gives classification results in excess of clinical standards. Using only time series data gives results comparable with existing clinical standards, given a long enough time series.

As current NICU patient data systems already collect sensory data used in this paper, predictive modeling could be included in the care process to give physicians advance warning of increased risk of in-hospital death. The model already outperforms existing methods in our retrospective cohort and with further refinement could prove to be a valuable clinical tool.

Acknowledgement. The authors would like to thank Aki Vehtari for help in Gaussian process classifier design and implementation, and the Academy of Finland (projects 295505 and 266940) for financial support. We acknowledge the computational resources provided by the Aalto Science-IT project.

References

[1] N. Byrnes, Can Technology Fix Medicine?, MIT Technology Review (Sep/Oct 2014).

[2] D. A. Clifton, K. E. Niehaus, P. Charlton, G. W. Colopy, Health Informatics via Machine Learning for the Clinical Management of Patients:, IMIA Yearbook 10 (1) (2015) 38–43. `doi:10.15265/IY-2015-014`.

[3] D. You, L. Hug, S. Ejdemyr, J. Beise, Levels and trends in child mortality. Report 2015. Estimates developed by the UN Inter-agency Group for Child

Mortality Estimation., Tech. rep., United Nations Inter-agency Group for Child Mortality Estimation (UN IGME), New York, NY (2015).

[4] R. M. Patel, S. Kandefer, M. C. Walsh, E. F. Bell, W. A. Carlo, A. R. Laptook, P. J. Sánchez, S. Shankaran, K. P. Van Meurs, M. B. Ball, E. C. Hale, N. S. Newman, A. Das, R. D. Higgins, B. J. Stoll, Causes and Timing of Death in Extremely Premature Infants from 2000 through 2011, New England Journal of Medicine 372 (4) (2015) 331–340. doi:10.1056/NEJMoa1403489.

[5] D. K. Richardson, J. E. Gray, M. C. McCormick, K. Workman, D. A. Goldmann, Score for Neonatal Acute Physiology: A physiologic severity index for neonatal intensive care, Pediatrics 91 (3) (1993) 617–623.

[6] D. K. Richardson, J. D. Corcoran, G. J. Escobar, S. K. Lee, SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores, The Journal of Pediatrics 138 (1) (2001) 92–100. doi:10.1067/mpd.2001.109608.

[7] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2006.

[8] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[9] J. C. Platt, Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, in: Advances in Large Margin Classifiers, MIT Press, 1999, pp. 61–74.

[10] M. Alvarez, D. Luengo, N. Lawrence, Latent force models, in: Artificial Intelligence and Statistics, 2009, pp. 9–16.

[11] M. A. Alvarez, D. Luengo, N. D. Lawrence, Linear Latent Force Models Using Gaussian Processes, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (11) (2013) 2693–2705. doi:10.1109/TPAMI.2013.86.

[12] J. Hartikainen, S. Särkkä, Sequential inference for latent force models, Proceedings of The 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011).

[13] J. Hartikainen, M. Seppänen, S. Särkkä, State-space inference for non-linear latent force models with application to satellite orbit prediction, in: Proceedings of the 29th International Conference on Machine Learning (ICML-12), Edinburgh, Scotland, UK, 2012, pp. 903–910.

[14] S. Särkkä, A. Solin, J. Hartikainen, Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing: A look at Gaussian process regression through Kalman filtering, IEEE Signal Processing Magazine 30 (4) (2013) 51–61.

[15] S. Särkkä, Bayesian Filtering and Smoothing, Cambridge University Press, 2013.

[16] G. W. Colopy, M. A. Pimentel, D. A. Clifton, S. J. Roberts, Bayesian Gaussian Processes for Identifying the Deteriorating Patient, in: Proceedings of the 38th Annual Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, Orlando, FL, USA, 16-20 Aug 2016, pp. 5311–5314. doi:10.1109/EMBC.2016.7591926.

[17] M. Ghassemi, M. A. Pimentel, T. Naumann, T. Brennan, D. A. Clifton, P. Szolovits, M. Feng, A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data, in: Proceedings of the 29th AAAI Conference on Artificial Intelligence, Vol. 1, NIH Public Access, Austin, TX, USA, 2015, pp. 446–453.

[18] F. Güiza, J. Ramon, H. Blockeel, Gaussian processes for prediction in intensive care, in: Gaussian Processes in Practice Workshop, Bletchley Park, UK, 2006, pp. 1–4.

[19] V. Gangadharan, Automated multi-parameter monitoring of neonates, PhD thesis, UCL (University College London), London (2013).

[20] S. Saria, A. K. Rajani, J. Gould, D. Koller, A. A. Penn, Integration of Early Physiological Responses Predicts Later Illness Severity in Preterm Infants, Science Translational Medicine 2 (48) (2010) 48ra65–48ra65. `doi:10.1126/scitranslmed.3001304`.

[21] S. Reid, B. Bajuk, K. Lui, E. A. Sullivan, NSW and ACT Neonatal Intensive Care Units Audit Group, PSN, Comparing CRIB-II and SNAPPE-II as mortality predictors for very preterm infants: Comparing CRIB-II and SNAPPE-II, Journal of Paediatrics and Child Health 51 (5) (2015) 524–528. `doi:10.1111/jpc.12742`.

[22] O.-P. Rinta-Koski, S. Särkkä, J. Hollmén, M. Leskinen, S. Andersson, Prediction of preterm infant mortality with Gaussian process classification, in: Proceedings of the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 26-28 April 2017, pp. 193–198.

[23] J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, A. Vehtari, GPstuff: Bayesian modeling with Gaussian processes, Journal of Machine Learning Research 14 (Apr) (2013) 1175–1179.

[24] A. Solin, S. Särkkä, The 10th annual MLSP competition: First place, in: 2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2014, pp. 1–3.

[25] K. Suotsalo, S. Särkkä, Detecting Malignant Ventricular Arrhythmias in Electrocardiograms by Gaussian Process Classification, in: Proceedings of the 27th IEEE International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, Tokyo, Japan, September 25-28, 2017.

[26] J. A. Swets, Measuring the Accuracy of Diagnostic Systems, Science 240 (1988) 1285–1293.

[27] A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern recognition 30 (7) (1997) 1145–1159.

[28] W. J. Youden, Index for rating diagnostic tests, Cancer 3 (1950) 32–35. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

[29] H. Rue, S. Martino, N. Chopin, Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (2) (2009) 319–392. doi:10.1111/j.1467-9868.2008.00700.x.

[30] C. Bührer, B. Metze, M. Obladen, CRIB, CRIB-II, birth weight or gestational age to assess mortality risk in very low birth weight infants?, Acta Paediatrica 97 (7) (2008) 899–903. doi:10.1111/j.1651-2227.2008.00793.x.

[31] The International Neonatal Network, The CRIB (clinical risk index for babies) score: A tool for assessing initial neonatal risk and comparing performance of neonatal intensive care units, The Lancet 342 (8865) (1993) 193–198.

[32] G. Parry, J. Tucker, W. Tarnow-Mordi, CRIB II: An update of the clinical risk index for babies score, The Lancet 361 (9371) (2003) 1789–1791.

[33] L. Gagliardi, A. Cavazza, A. Brunelli, M. Battaglioli, D. Merazzi, F. Tandoi, D. Cella, G. F. Perotti, M. Pelti, I. Stucchi, F. Frisone, A. Avanzini, R. Bellù, the NNL study group, Assessing mortality risk in very low birthweight infants: A comparison of CRIB, CRIB-II, and SNAPPE-II, Archives of Disease in Childhood - Fetal and Neonatal Edition 89 (5) (2004) F419–F422. doi:10.1136/adc.2003.031286.