Jaskari, Joel; Myllarinen, Janne; Leskinen, Markus; Rad, Ali Bahrami; Hollmén, Jaakko; Andersson, Sture; Sarkka, Simo

# Machine Learning Methods for Neonatal Mortality and Morbidity Classification

*Published in:*
IEEE Access

*DOI:*
10.1109/ACCESS.2020.3006710

Published: 02/07/2020

*Document Version*
Publisher's PDF, also known as Version of record

Received June 10, 2020, accepted June 20, 2020, date of publication July 2, 2020, date of current version July 17, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3006710*

# Machine Learning Methods for Neonatal Mortality and Morbidity Classification

**JOEL JASKARI[1], JANNE MYLLÄRINEN[2], MARKUS LESKINEN[3], ALI BAHRAMI RAD[2,4], (Member, IEEE), JAAKKO HOLLMÉN[1,5], (Senior Member, IEEE), STURE ANDERSSON[3], AND SIMO SÄRKKÄ[2], (Senior Member, IEEE)**

[1]Department of Computer Science, Aalto University, 00076 Aalto, Finland
[2]Department of Electrical Engineering and Automation, Aalto University, 00076 Aalto, Finland
[3]Pediatric Research Center, Children's Hospital, University of Helsinki, Helsinki University Hospital, 00029 Helsinki, Finland
[4]Department of Biomedical Informatics, Emory University, Atlanta, GA 30322, USA
[5]Department of Computer and Systems Sciences, Stockholm University, 16407 Stockholm, Sweden

Corresponding author: Simo Särkkä (simo.sarkka@aalto.fi)

The work of Sture Andersson was supported by the Foundation for Pediatric Research in Finland, and a Special Governmental Subsidy for Clinical Research.

**ABSTRACT** Preterm birth is the leading cause of mortality in children under the age of five. In particular, low birth weight and low gestational age are associated with an increased risk of mortality. Preterm birth also increases the risks of several complications, which can increase the risk of death, or cause long-term morbidities with both individual and societal impacts. In this work, we use machine learning for prediction of neonatal mortality as well as neonatal morbidities of bronchopulmonary dysplasia, necrotizing enterocolitis, and retinopathy of prematurity, among very low birth weight infants. Our predictors include time series data and clinical variables collected at the neonatal intensive care unit of Children's Hospital, Helsinki University Hospital. We examine 9 different classifiers and present our main results in AUROC, similar to our previous studies, and in F1-score, which we propose for classifier selection in this study. We also investigate how the predictive performance of the classifiers evolves as the length of time series is increased, and examine the relative importance of different features using the random forest classifier, which we found to generally perform the best in all tasks. Our systematic study also involves different data preprocessing methods which can be used to improve classifier sensitivities. Our best classifier AUROC is 0.922 in the prediction of mortality, 0.899 in the prediction of bronchopulmonary dysplasia, 0.806 in the prediction of necrotizing enterocolitis, and 0.846 in the prediction of retinopathy of prematurity. Our best classifier F1-score is 0.493 in the prediction of mortality, 0.704 in the prediction of bronchopulmonary dysplasia, 0.215 in the prediction of necrotizing enterocolitis, and 0.368 in the prediction of retinopathy of prematurity.

**INDEX TERMS** Bronchopulmonary dysplasia, classification, machine learning, necrotizing enterocolitis, neonatal intensive care unit, neonatal mortality, neonatology, NICU, retinopathy of prematurity.

## I. INTRODUCTION

Over 15 million babies are born preterm every year, and while their mortality and morbidity rates have been decreasing in recent decades, preterm birth is still the worldwide leading cause of childhood mortality under the age of five [1]. Increased risk of mortality and morbidity among neonates is associated with low birth weight and low gestational age [2]. Very low birth weight (VLBW) infants, that is, those with birth weight under 1500 g, which are treated in neonatal intensive care units (NICUs) in the Western Europe and in the USA, have a mortality rate around 11% [3]. Furthermore, many of the survivors develop severe complications such as neonatal sepsis [4], bronchopulmonary dysplasia (BPD) [5], or necrotizing enterocolitis (NEC) [6]. These and other complications can also inflict long-term or permanent morbidities such as persistent pulmonary dysfunction in the case of BPD [5], gastrointestinal and neurodevelopmental problems in the case of NEC [6], or blindness in the case of retinopathy of prematurity (ROP) [7]. Early detection of neonatal morbidities is of paramount importance to halt the progression of the disease, and for preventing further complications or even death [8].

In order to better assess the risk of mortality and neonatal illness, several scores have been proposed, such as the Apgar [9], SNAP [10], and SNAPPE [11] scores, and more

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Li.

recently the SNAP-II and SNAPPE-II [12] scores. Modern NICUs monitor the vital signs of neonates in an automated fashion, which allows for the collection of time series data of the physiological measurements. In our preliminary work, we observed that machine learning methods can enhance the performance of traditional SNAP-II and SNAPPE-II scores by leveraging time series data. The methodology leads to improved prediction of neonatal mortality [13] and neonatal morbidities of BPD, NEC, and ROP [14]. In this study, we continue the mentioned preliminary work by presenting a comprehensive comparison of machine learning methods for prediction of mortality, BPD, NEC, and ROP in VLBW neonates.

The contributions of this article are i) to present a systematic study of machine learning methods for the prediction of neonatal mortality, BPD, NEC, and ROP, ii) to propose and analyze F1-score as an evaluation measure to AUROC in order to improve the sensitivity of the classifiers, iii) to analyze the methods using various lengths of time series data, and iv) to analyze the relative importance of different features using the random forest classifier, which we found to have generally the best performance in the classification tasks. This study was approved by the Ethics Committee of Helsinki University Hospital 115/13/03/00/14, dated April 8, 2014.

## II. BACKGROUND

Bronchopulmonary dysplasia (BPD) is a severe chronic lung complication among the preterm infants resulting from the immaturity of the developing preterm lung and injuries associated with external effects, such as maternal intra-amniotic infection, mechanical ventilation, and excessive oxygen [15]. The initial injury is often caused by respiratory distress syndrome (RDS) [15] or acute respiratory distress [5]. However, oxygen and positive-pressure ventilation system used for treating these conditions worsens the injury and initiates the development of BPD [15].

Necrotizing enterocolitis (NEC) is a serious disease of the developing gastrointestinal tract. Early diagnosis of NEC is important in order to prevent the inflammation to progress to bowel necrosis and perforation, and eventually to death [8]. There is also an economic aspect, as it has been estimated that the annual costs of the disease are between $500 million and $1 billion in the USA [16].

Retinopathy of prematurity (ROP) is a retinal disease associated with the combination of underdeveloped retinal vessels of preterm neonates and supplementary oxygen given to them [17]. In addition, low gestational age and birth weight are known risk factors. Currently, the screening for ROP is performed by eye examinations and treated with laser [7] or with anti-vascular endothelial growth factor treatment [18].

SNAP-II and SNAPPE-II [12] are scores for predicting neonatal mortality and they are based on a simple logistic regression model. SNAP-II uses features extracted from the physiological measurements and laboratory results, such as mean blood pressure and lowest serum pH, during a 12 hour recording period. SNAPPE-II score additionally uses the birth weight, the gestational age, and the Apgar score. The Apgar score is reported at 1 minute and 5 minutes after birth for all infants, and at 5-minute intervals thereafter until 20 minutes for infants with a score less than 7, to assess the status of the infant and the response to resuscitation if needed. The Apgar score is a sum of five indices measuring the heart rate, respiratory effort, muscle tone, reflex irritability, and color of the infant [19].

Using machine learning for neonatal morbidity and mortality prediction has been previously explored in literature. Saria *et al.* [20] presented a model for predicting binary label of low/high morbidity where, high morbidity was defined as any of the following complications: sepsis, pulmonary hemorrhage, pulmonary hypertension, acute hemodynamic instability, moderate or severe BPD, ROP, NEC, intraventricular hemorrhage, or death. Their model utilized aggregated nonlinear Bayesian models and logistic regression, and their selected features included mean values, baseline variances, and residual variances of 3 hours of recorded heart rate, respiratory rate, and oxygen saturation. In addition, the gestational age and birth weight were used as inputs to the model. Our selected features are similar, as we use the mean and standard deviation of the heart rate and oxygen saturation, the gestational age, and the birth weight. However, our selected signals differ in that we do not use respiratory rate and we use additional features including systolic, diastolic, and mean blood pressure, as well as the SNAP-II and SNAPPE-II scores. We also consider a larger variety of popular classifiers for tabular data. Saria *et al.* [20] also considered the importance of physiological features. However, we analyze the features from the perspective of random forest feature importance, while in Saria *et al.* the importance was estimated using ablation analysis. We have also chosen to predict mortality and different morbidities separately as opposed to the combined high morbidity label used in the article.

In a more recent work, Podda *et al.* [21] presented multiple machine learning methods for predicting neonatal mortality in the terms of the probability of survival: logistic regression, $k$-nearest neighbor, random forest, gradient boosting machine, support vector machine, and neural network. Our work also considers logistic regression, $k$-nearest neighbor, random forest, and support vector machine, but we also use five additional classifiers and evaluate our classifiers in predicting different morbidities. In addition, our features differ from the ones used in Podda *et al.* [21]. The models presented in Podda *et al.* [21] do not use physiological time series, but instead, they use the following data collected up to 5 minutes after birth: gestational age, birth weight, Apgar scores 1 minute and 5 minutes after the birth, sex, multiple gestation, mode of delivery, prenatal care, intra-amniotic infection, maternal hypertension, ethnicity, and antenatal steroids. Our dataset includes data up to 72 hours, from which we use multiple subsets in order to examine how the predictive performance of the models vary with the time series length.

Our work is a continuation of Rinta-Koski *et al.* [14] and Rinta-Koski *et al.* [13]. These works presented preliminary

**TABLE 1. Descriptive statistics of the set of included patients.**

| | All | In-Hospital Death | BPD | NEC | ROP |
|---|---|---|---|---|---|
| VLBW infants (n) | 977 | 63 | 275 | 31 | 77 |
| Gestational age - days (mean,std) | 196.4±14.3 | 178.9±12.1 | 184.8±10.6 | 182.2±11.4 | 180.8±9.1 |
| Birth weight - grams (mean,std) | 1037±263 | 678±186 | 852±205 | 817±261 | 785±187 |
| Proportion of Male infants (%) | 50.5 | 66.7 | 56.7 | 71.0 | 63.6 |
| Proportion of Female infants (%) | 49.5 | 33.3 | 43.3 | 29.0 | 36.4 |
| Time in NICU - days (mean,std) | 29.7±27.0 | 23.6±34.5 | 57.0±27.2 | 58.4±36.4 | 78.4±29.1 |

results obtained for a small number of classifiers for mortality, BPD, NEC, and ROP prediction on a dataset collected in the NICU of Children's Hospital of Helsinki University Hospital. In Rinta-Koski *et al.* [14], Gaussian process classifiers were trained to predict BPD, NEC, and ROP, and the results were compared to the standard medical scores SNAP-II and SNAPPE-II. In Rinta-Koski *et al.* [13], neonatal mortality prediction was considered. The classifiers compared in the study were Gaussian process classifier, with three different kernels, support vector machine classifier, linear probit model, SNAP-II, and SNAPPE-II.

## III. DATA

Our dataset consists of pseudonymized temporal and static data collected in the NICU of Children's Hospital, Helsinki University Hospital between years 1999 and 2013. This dataset is partially the same dataset used in [14] and [13]. However, in this study, we have access to additional patients. We included all the infants admitted to NICU with birth weights under 1500 g (VLBW) in this study. Patients who died or were discharged before the age of 72 hours were excluded in order to prevent the explicit vital sign decay to affect the predictions. Also, patients with less than 50 measurements of any of the time series predictors were excluded. Patients with severe congenital anomalies were not excluded, as we wanted an inclusive sample and birth defects are an important cause of neonatal morbidity and mortality among the VLBW infants. The study included 977 patients, who fulfilled the inclusion criteria. Descriptive statistics of the set of included patients are presented in Table 1.

Temporal data includes physiological variables, such as heart rate and blood pressure, in the form of time series. Static data includes clinical information, such as gestational age and birth weight, medical scores SNAP-II and SNAPPE-II, diagnoses of the patients for BPD, NEC, and ROP, as well as information on the survival. There are up to 111 different sensor measurements, however, many, or even all measurements were missing for a large portion of the patients, and thus only a subset of the features was selected. The predictors were chosen to be the same as in [13]: systolic, diastolic and mean blood pressure, oxygen saturation and heart rate for temporal variables, and for static variables SNAP-II, SNAPPE-II, birth weight, and gestational age at birth.

We use these predictors because our preliminary work shows that the best result can be obtained when all these predictors are used in the classification. It is to be noted, that there is slight overlap in variables, as SNAP-II and SNAPPE-II scores also include information about the mean blood pressure, oxygen saturation, and birth weight. Relationship between SNAP-II and SNAPPE-II is also very close. However, the information in SNAP-II and SNAPPE-II is nonlinear in nature, due to the thresholding of the scores, and thus it is not redundant to include the constituent features. Using SNAP-II and SNAPPE-II is in a sense inclusion of prior medical knowledge on how the constituent variables should be processed. Analysing the individual constituents of SNAP-II and SNAPPE-II would be beneficial, as these scores might not be computed in every hospital, making the approach more general. This analysis is, however, left for future work.

The temporal variables were collected using sensor measurements throughout the stay of the patient in the NICU. The gestational age was expressed as completed days from mother's last menstrual period to delivery. In our clinical practice, voluntary early ultrasonographic examination is offered to all mothers, which is used for estimating the expected gestational age. If the difference between the gestational age, calculated from last menstruation period, and the expected gestational age is more than five days, gestational age is adjusted.

The sampling of the temporal variables is irregular in the hospital environment, meaning that the sampling of the sensor measurements is not synchronized. Our dataset also has a large number of missing values. We applied multiple preprocessing steps in order to standardize the data into a format suitable for machine learning methods, and also to examine the behavior of these algorithms in various circumstances. The non-synchronous data was transformed into another set, by splitting the time into 2-minute intervals and by filling missing data with the closest observation before the start of the interval, thus creating a synthetic synchronous dataset.

VLBW infants undergo an adaptation period after birth when their physiological recordings differ from the steady state attained later. We hypothesized that these initial adaptation period recordings might affect the classifier performance. As the age, when VLBW infants were admitted to NICU varied, we created two new sets from both the non-synchronous and synchronous datasets with the first 6 hours of measurements excluded, in order to examine if the more

stable post-adaptation period recordings improve the performance of the classifiers.

At the end, we created multiple subsets from the previous 4 sets, by dividing the time series into 12, 18, 24, 36, 48, and 72 hours of data. Further exclusion of patients was conducted on the basis of observed sensor measurements. Measurements out of range of plausible values, defined by medical experts, were removed, and, again, at least 50 valid measurements per temporal variable were required to include the patient into the dataset. This minimum 50-valid-measurements rule causes different sets to have a different number of patients.

Detailed descriptions of the datasets produced by the different preprocessing steps are presented in the Supplementary material in Table 1a for mortality, Table 1b for BPD, Table 1c for NEC, and Table 1d for ROP. From the patients in our dataset, approximately 6-7% died, 28-29% had BPD, 3% NEC, and 8% ROP.

We also experimented with undersampling of the majority class during the training of the classifiers to overcome the class imbalance problem. This procedure is explained in the next section. Lastly, our preprocessing pipeline includes feature extraction and data standardization. Similarly to [13] and [14], we chose to use the means and standard deviations as our extracted features from the time series. We also normalized the data to have zero mean and unit variance, for each variable independently, using the mean and the standard deviation of the training set of each fold in order to prevent leakage between the sets.

## IV. CLASSIFIER METHODS

Our selected methods include logistic regression, linear discriminant analysis, quadratic discriminant analysis, $k$-nearest neighbor, support vector machine, three different Gaussian processes, and random forest classifier.

In logistic regression (LR), a linear model is combined with logistic sigmoid function to model the posterior probabilities of each class [22]. The resulting model is nonlinear and there is no closed-form solution, hence iterative optimization methods like Newton-Raphson [23] or gradient descent [24] have to be used. The probabilities of the classes can be defined for a binary classification task as is presented in Equation (1) [23], bias term included in $\boldsymbol{w}$ for clarity:

$$\begin{aligned} p(y = 1 \mid \boldsymbol{x}, \boldsymbol{w}) &= \frac{1}{1 + \exp(-\boldsymbol{w}^T \boldsymbol{x})}, \\ p(y = 0 \mid \boldsymbol{x}, \boldsymbol{w}) &= 1 - p(y = 1 \mid \boldsymbol{x}, \boldsymbol{w}). \end{aligned} \tag{1}$$

Linear discriminant analysis (LDA) classifier [25] is based on the assumption that the class-conditional densities $p(\boldsymbol{x} \mid y = 1)$ and $p(\boldsymbol{x} \mid y = 0)$ are Gaussian. In addition, LDA further assumes that the density functions share a common covariance matrix. Under these assumptions, the classification problem can be formulated as finding the optimal 1D projection for the data, governed by the class-conditional means and common covariance matrix, and finding the optimal classification threshold on this line. The equations for the optimal projection vector and the classification threshold on the projection, given uniform class priors, are presented in Equations (2) and (3), respectively, for binary case [25]:

$$\boldsymbol{w} = \Sigma^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0}), \tag{2}$$

$$\Xi = \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})^T \Sigma^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0}). \tag{3}$$

In these equations, $\boldsymbol{\mu}_n$ denotes the mean of the samples from class $n$, and $\Sigma$ is the common covariance matrix.

Quadratic discriminant analysis (QDA) classifier [25] is a generalization of the LDA classifier, where the requirement of the common covariance matrix is relaxed into class specific covariance matrices. This relaxation also causes the decision boundaries between classes to have a quadratic form. The posterior probabilities of classes can be computed as in Equation (4), for uniform priors again for clarity [25]:

$$p(y = c \mid \boldsymbol{x}) = \frac{|\Sigma_c|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_c)^T \Sigma_c^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_c)\}}{\sum_{i \in C} |\Sigma_i|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\}}. \tag{4}$$

In this equation, $i$ denotes the index of the class, $c$ of which is the predicted class, and $\boldsymbol{\mu}_i$ and $\Sigma_i$ are the mean vector and covariance matrix of the class $i$, respectively.

$k$-nearest neighbor (KNN) classification algorithm is a machine learning method, using a distance metric, such as Euclidean distance, to find the closest examples in the training set to a query point, and then classifying points based on these neighboring points [22]. The posterior probability of a query point belonging to a specific class can be estimated as the proportion of neighbors belonging to that class. The class associated with the highest posterior probability can then be used as the prediction of the model [23].

Support vector machine (SVM) [22] is based on finding a hyperplane that separates the samples between the classes with maximal margin. If the classes are not linearly separable, so-called slack variables are needed for each data point, such that some may fall on the wrong side of the separating hyperplane. The slack variables are constrained to be non-negative, and the separating hyperplane is selected such that the sum of the slack variables is minimized. So-called kernel trick can be used to lift the problem into a higher-dimensional space [22]. However, we do not consider the kernel SVMs in this work.

Gaussian processes (GP) [26] are Bayesian non-parametric machine learning methods, which can be used in classification and regression tasks. Gaussian process defines a Gaussian distribution over functions and it is parametrized by a mean function and a covariance function. In our experiments we use zero mean, and thus the covariance function defines our GP model.

In this work, three covariance functions have been utilized, each consisting of a sum of constant kernel, linear kernel, and one of the following: squared exponential kernel ($k_{\text{RBF}}(\cdot,\cdot)$), Matérn kernel with $\nu = 3/2$ ($k_{\text{Matérn32}}(\cdot,\cdot)$), or Matérn kernel with $\nu = 5/2$ ($k_{\text{Matérn52}}(\cdot,\cdot)$). The three latter kernels are presented in Equations (5)-(7) with parameter $\nu$ substituted

with the relevant value:

$$k_{\text{RBF}}(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^2), \tag{5}$$

$$k_{\text{Matérn32}}(\mathbf{x}, \mathbf{x}') = (1 + \frac{\sqrt{3}|\mathbf{x} - \mathbf{x}'|}{l}) \cdot \exp(-\frac{\sqrt{3}|\mathbf{x} - \mathbf{x}'|}{l}), \tag{6}$$

$$k_{\text{Matérn52}}(\mathbf{x}, \mathbf{x}') = (1 + \frac{\sqrt{5}|\mathbf{x} - \mathbf{x}'|}{l} + \frac{5|\mathbf{x} - \mathbf{x}'|^2}{3l^2}) \cdot \exp(-\frac{\sqrt{5}|\mathbf{x} - \mathbf{x}'|}{l}). \tag{7}$$

Bootstrap aggregation of multiple decision trees gives arise to a model called random forest (RF) [22]. In RF, each tree is given a random bootstrap sample of the training dataset for which splitting rules are learned. RF algorithm also includes random sampling of the features at each split, in order to reduce the correlation between the trained trees. The predictions of RF are obtained by aggregating all the tree-predictions together by averaging them [22].

## V. EVALUATION CRITERIA

We evaluate our models by using two primary evaluation measures: the area under the receiver operating characteristics curve (AUROC) and the F1-score, also called the Sørensen-Dice coefficient and the similarity coefficient. AUROC [27] is a common measure used in machine learning method evaluation within the medical field, see for example [13], [14], [28], [29], and is therefore selected for evaluation. However, our dataset has a high class imbalance, and thus in this study, we propose the F1-score as an alternative evaluation measure to better reflect the classifier performance in detecting the positive class, that is, the minority class in our case.

As an example of using the AUROC measure for classifier selection with high class imbalance, the best classifier in the detection of mortality in our previous study [13] had a high AUROC of 0.948, however, the sensitivity was 0.463. This means that nearly 54% of the deaths are undetected by the classifier. Similarly in the detection of ROP in our previous study [14], the best classifier had a moderately high AUROC of 0.84 and a very low sensitivity of 0.05, which results in 95% of the ROP cases to be left undetected. We will show (see Section VI), when we have a class imbalance problem, selecting the best classifier based on the highest F1-score will significantly improve the sensitivity with possibly only a slight decrease in the AUROC value. The F1-score is computed as the harmonic mean of positive predictive value (PPV/precision) and sensitivity (recall) on a single operating point.

We also present results on the mean F1-score, defined as the mean of F1-score calculated for detection of positive class and for the detection of negative class, PPV, sensitivity, specificity, and accuracy. The equations for calculating the evaluation measures and their relevant intermediate results, other than AUROC, are the following Equations (8)-(15):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{8}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{9}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{10}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}, \tag{11}$$

$$\text{F1} = 2\,\frac{\text{PPV} \times \text{Sensitivity}}{\text{PPV} + \text{Sensitivity}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{12}$$

$$\text{F1}^- = 2\,\frac{\text{NPV} \times \text{Specificity}}{\text{NPV} + \text{Specificity}} = \frac{2\text{TN}}{2\text{TN} + \text{FP} + \text{FN}}, \tag{13}$$

$$\text{Mean F1} = \frac{\text{F1} + \text{F1}^-}{2}, \tag{14}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{15}$$

Here TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. NPV is abbreviation for negative predictive value.

Due to the small number of samples in our dataset, cross-validation was used. We used 8-fold cross-validation, repeated 8 times, similarly as in our preliminary work [13]. Stratification was also used in the cross-validation, which ensures that a similar proportion of positive and negative samples are selected into each fold. Due to the high class imbalance present in our dataset, we performed additional experiments with class sub-sampling in our experiments. In this setting, we modified the training set of each cross-validation fold to have the same amount of negative and positive examples by sampling the majority class without replacement. Predictions were performed on the entire test fold.

Parameter selection was conducted using a nested cross-validation procedure, where the standard cross-validation training set was used in another inner cross-validation loop. The validation set of this inner loop was then used to estimate the generalization performance with the selected parameters. With this procedure we searched for the best parameters for RF and KNN. For RF these parameters were the number of randomly selected variables for each split, number of trees within the forest, and minimum number of observations for each leaf node. For KNN, the parameter was the number of neighbors. These parameters are presented in Table 2.

We used MATLAB's inbuilt functions to implement our experiments, except for the Gaussian process classifier, for which we used the GPstuff MATLAB toolbox [30]. Evaluation measures were calculated utilizing Python framework Scikit-learn [31] for each model.

**TABLE 2. Selected parameters for random forest and KNN (*k*-nearest neighbor) for each classification task. PPS denotes predictors per split and OPL observations per leaf.**

| | Random Forest | | | KNN |
|---|---|---|---|---|
| | Trees | No. PPS | No. OPL | Neighbors |
| Mortality | 600 | 3 | 8 | 16 |
| BPD | 600 | 2 | 2 | 11 |
| NEC | 50 | 2 | 13 | 10 |
| ROP | 600 | 1 | 12 | 14 |

## VI. RESULTS

### A. CLASSIFICATION EXPERIMENTS

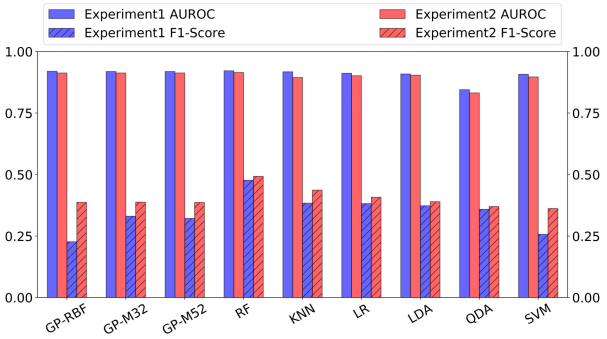
In this section, we present the results of our experiments. We first present results with classifier selection based on the AUROC, similar manner to our previous studies. We continue by presenting results with classifier selection based on the F1-score and show how classifiers selected based on this measure relate to the classifiers selected with AUROC. Full classification results for classifiers selected based on the best AUROC are presented in Table 3, and for classifiers selected based on the best F1-score are shown in Table 4. Full graphical comparison between these experiments is presented in Figure 1.

In mortality classification (Table 3a), all classifiers except the QDA achieved over 0.9 AUROC. The RF classifier had the highest AUROC of 0.922, however, with only a small margin to the Gaussian process classifiers with the GP-RBF having AUROC of 0.920, GP-M32 with 0.919, GP-M52 with 0.919, and to the KNN with AUROC of 0.918. The RF had also the highest F1-score of 0.477, in which the margin was larger, as the second best F1-score of 0.384 was achieved by the KNN. Graphical illustration of the results is presented in Figure 1a.

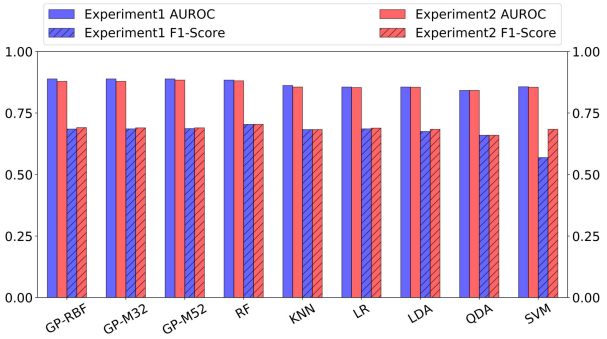
In BPD classification (Table 3b), no classifier achieved over 0.9 AUROC, the highest being the Gaussian process classifiers with 0.899 AUROC, however, only a small margin from the RF with 0.884 AUROC. A small performance gap can be seen in the F1-score, as the RF had the highest F1-score of 0.704 and the second highest F1-score was achieved by the GP-M52 classifier with 0.687 F1-score. All the Gaussian process classifiers had virtually the same performance in this task. The logistic regression classifier had a similar F1-score as the Gaussian process classifiers with 0.686 value. Graphical illustration of the results is presented in Figure 1b.

In NEC classification (Table 3c), the RF had the best AUROC of 0.806, which was the only result over 0.8. The RF also had the best F1-score of 0.189. Interestingly, all the Gaussian process classifiers and the SVM classifier failed to detect any positive examples. Also, all the other classifiers except for the RF, Gaussian process, and SVM had over 0.6 sensitivity and less than 0.1 PPV, indicating high amount of false positives. Graphical illustration of the results is presented in Figure 1c.

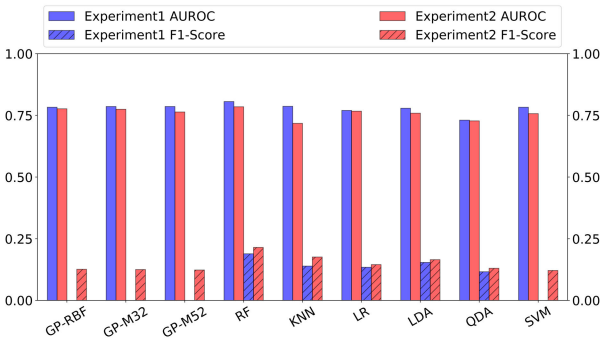
In ROP classification (Table 3d), the Gaussian process classifiers and the RF classifier had all the best AUROC of 0.846. However, the RF had slightly less variation in AUROC between the cross-validation folds and repetitions compared to the Gaussian process classifiers, indicated by



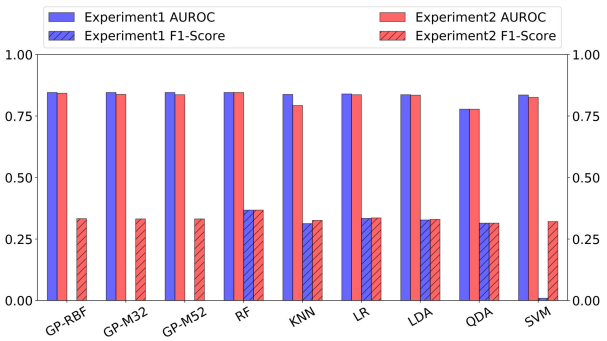

(a) Mortality prediction.




(b) BPD prediction.




(c) NEC prediction.




(d) ROP prediction.

**FIGURE 1. Visualization of our results as bar graphs. Experiment1 denotes the first experiment, where we selected the classifier for maximal AUROC, and the Experiment2 denotes the second experiment, where we selected the classifier for maximal F1-Score.**

**TABLE 3. Classification results for representative classifier selected based on AUROC. The results are presented as the mean and in parentheses the standard error of each evaluation measure, over the cross-validation repetitions. Descriptors under the name of the method are in format: length of time series (hours) - irregularly or regularly sampled (I/R) - all data or exclude first 6 hours (A/E) - class distribution empirical or resampled to uniform (E/U). GP denotes the Gaussian process classifier, with RBF denoting the squared exponential kernel, M32 the Matérn kernel with $\nu = 3/2$, and M52 the Matérn kernel with $\nu = 5/2$. RF denotes the random forest classifier, KNN the $k$-nearest neighbor classifier, LR the logistic regression classifier, LDA the linear discriminant analysis classifier, QDA the quadratic discriminant analysis classifier, and SVM the support vector machine classifier.**

| Method | AUROC | F1-Score | Mean F1-Score | PPV | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| GP-RBF 18h-I-A-E | 0.920 (0.02) | 0.227 (0.05) | 0.596 (0.03) | 0.493 (0.12) | 0.157 (0.04) | 0.988 (0.00) | 0.934 (0.00) |
| GP-M32 72h-I-A-E | 0.919 (0.02) | 0.331 (0.06) | 0.650 (0.03) | **0.624** (0.10) | 0.242 (0.05) | 0.989 (0.00) | **0.941** (0.00) |
| GP-M52 36h-R-E-E | 0.919 (0.02) | 0.322 (0.06) | 0.645 (0.03) | 0.586 (0.11) | 0.240 (0.05) | 0.987 (0.00) | 0.938 (0.01) |
| RF 36h-I-A-E | **0.922** (0.01) | **0.477** (0.04) | **0.712** (0.02) | 0.377 (0.03) | 0.674 (0.05) | 0.919 (0.01) | 0.903 (0.01) |
| KNN 72h-R-A-U | 0.918 (0.01) | 0.384 (0.02) | 0.634 (0.01) | 0.245 (0.01) | **0.906** (0.03) | 0.800 (0.02) | 0.807 (0.01) |
| LR 72h-I-E-U | 0.912 (0.01) | 0.382 (0.02) | 0.638 (0.01) | 0.248 (0.02) | 0.854 (0.04) | 0.818 (0.01) | 0.820 (0.01) |
| LDA 18h-I-A-E | 0.909 (0.02) | 0.373 (0.02) | 0.628 (0.02) | 0.238 (0.02) | 0.885 (0.04) | 0.796 (0.01) | 0.802 (0.01) |
| QDA 36h-R-A-E | 0.845 (0.02) | 0.359 (0.03) | 0.634 (0.02) | 0.248 (0.02) | 0.674 (0.05) | 0.852 (0.01) | 0.840 (0.01) |
| SVM 72h-R-A-E | 0.908 (0.01) | 0.257 (0.07) | 0.612 (0.03) | 0.529 (0.14) | 0.180 (0.05) | **0.990** (0.00) | 0.938 (0.00) |

(a) Mortality

| Method | AUROC | F1-Score | Mean F1-Score | PPV | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| GP-RBF 72h-R-A-E | **0.889** (0.01) | 0.685 (0.02) | 0.781 (0.01) | 0.690 (0.02) | 0.684 (0.03) | 0.877 (0.01) | 0.822 (0.01) |
| GP-M32 72h-R-A-E | **0.889** (0.01) | 0.686 (0.02) | 0.782 (0.01) | **0.698** (0.02) | 0.679 (0.03) | 0.882 (0.01) | **0.825** (0.01) |
| GP-M52 72h-R-A-E | **0.889** (0.01) | 0.687 (0.02) | 0.783 (0.02) | 0.697 (0.02) | 0.681 (0.03) | 0.882 (0.01) | **0.825** (0.01) |
| RF 72h-R-A-E | 0.884 (0.01) | **0.704** (0.02) | **0.785** (0.02) | 0.650 (0.02) | 0.771 (0.03) | 0.833 (0.01) | 0.815 (0.01) |
| KNN 72h-R-A-E | 0.862 (0.01) | 0.683 (0.02) | 0.761 (0.01) | 0.595 (0.02) | **0.806** (0.02) | 0.780 (0.01) | 0.787 (0.01) |
| LR 72h-R-E-E | 0.856 (0.01) | 0.686 (0.02) | 0.764 (0.01) | 0.602 (0.02) | 0.802 (0.03) | 0.787 (0.02) | 0.791 (0.01) |
| LDA 72h-R-E-E | 0.856 (0.01) | 0.675 (0.02) | 0.756 (0.01) | 0.592 (0.02) | 0.791 (0.03) | 0.781 (0.02) | 0.784 (0.01) |
| QDA 72h-R-E-E | 0.842 (0.01) | 0.660 (0.02) | 0.749 (0.01) | 0.595 (0.02) | 0.747 (0.03) | 0.794 (0.02) | 0.781 (0.01) |
| SVM 72h-I-E-E | 0.857 (0.01) | 0.569 (0.02) | 0.713 (0.01) | 0.655 (0.02) | 0.508 (0.03) | **0.894** (0.01) | 0.785 (0.01) |

(b) BPD

| Method | AUROC | F1-Score | Mean F1-Score | PPV | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| GP-RBF 72h-R-A-E | 0.783 (0.04) | 0.000 (0.00) | 0.492 (0.00) | 0.000 (0.00) | 0.000 (0.00) | **1.000** (0.00) | 0.968 (0.00) |
| GP-M32 72h-R-A-E | 0.786 (0.04) | 0.000 (0.00) | 0.492 (0.00) | 0.000 (0.00) | 0.000 (0.00) | **1.000** (0.00) | 0.968 (0.00) |
| GP-M52 72h-R-A-E | 0.786 (0.04) | 0.000 (0.00) | 0.492 (0.00) | 0.000 (0.00) | 0.000 (0.00) | **1.000** (0.00) | 0.968 (0.00) |
| RF 48h-R-A-E | **0.806** (0.03) | **0.189** (0.05) | **0.578** (0.03) | **0.163** (0.05) | 0.247 (0.07) | 0.960 (0.01) | 0.937 (0.01) |
| KNN 72h-I-A-U | 0.787 (0.04) | 0.139 (0.02) | 0.494 (0.01) | 0.078 (0.01) | 0.659 (0.09) | 0.746 (0.02) | 0.743 (0.02) |
| LR 72h-I-A-E | 0.770 (0.04) | 0.134 (0.02) | 0.490 (0.01) | 0.075 (0.01) | 0.642 (0.09) | 0.743 (0.01) | 0.739 (0.01) |
| LDA 72h-R-A-E | 0.779 (0.05) | 0.154 (0.02) | 0.509 (0.01) | 0.087 (0.01) | 0.664 (0.08) | 0.770 (0.01) | 0.766 (0.01) |
| QDA 72h-I-A-U | 0.731 (0.05) | 0.116 (0.01) | 0.448 (0.01) | 0.063 (0.01) | **0.711** (0.09) | 0.650 (0.03) | 0.652 (0.03) |
| SVM 72h-R-E-E | 0.783 (0.04) | 0.000 (0.00) | 0.492 (0.00) | 0.000 (0.00) | 0.000 (0.00) | **1.000** (0.00) | **0.969** (0.00) |

(c) NEC

| Method | AUROC | F1-Score | Mean F1-Score | PPV | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| GP-RBF 72h-I-E-E | **0.846** (0.02) | 0.000 (0.00) | 0.479 (0.00) | 0.000 (0.00) | 0.000 (0.00) | **1.000** (0.00) | **0.921** (0.00) |
| GP-M32 72h-R-A-E | **0.846** (0.02) | 0.000 (0.00) | 0.479 (0.00) | 0.000 (0.00) | 0.000 (0.00) | **1.000** (0.00) | 0.920 (0.00) |
| GP-M52 72h-R-A-E | **0.846** (0.02) | 0.000 (0.00) | 0.479 (0.00) | 0.000 (0.00) | 0.000 (0.00) | **1.000** (0.00) | 0.920 (0.00) |
| RF 72h-R-E-E | **0.846** (0.01) | **0.368** (0.03) | **0.632** (0.02) | **0.261** (0.02) | 0.643 (0.05) | 0.838 (0.01) | 0.822 (0.01) |
| KNN 72h-I-E-U | 0.838 (0.02) | 0.313 (0.02) | 0.560 (0.01) | 0.192 (0.01) | **0.854** (0.04) | 0.686 (0.02) | 0.700 (0.02) |
| LR 72h-I-E-E | 0.840 (0.02) | 0.334 (0.02) | 0.594 (0.01) | 0.215 (0.02) | 0.756 (0.05) | 0.760 (0.01) | 0.760 (0.01) |
| LDA 72h-I-E-E | 0.837 (0.02) | 0.328 (0.02) | 0.585 (0.02) | 0.209 (0.02) | 0.774 (0.05) | 0.743 (0.01) | 0.745 (0.01) |
| QDA 72h-I-E-E | 0.778 (0.02) | 0.315 (0.02) | 0.588 (0.01) | 0.208 (0.01) | 0.669 (0.05) | 0.777 (0.01) | 0.769 (0.01) |
| SVM 72h-I-E-E | 0.836 (0.02) | 0.009 (0.01) | 0.483 (0.01) | 0.039 (0.06) | 0.005 (0.01) | 0.997 (0.00) | 0.919 (0.00) |

(d) ROP

the 0.01 lower standard error. Similar to the NEC prediction, the Gaussian process classifiers failed to detect any positive examples and also the SVM had difficulties in detecting positive examples, reflected by the low sensitivity (0.005) and F1-score (0.009). On the other hand, the RF achieved the highest F1-score of 0.368 in addition to having the highest AUROC. Graphical illustration of the results is presented in Figure 1d.

In our second classification experiment we used the F1-score for classifier selection. The complete results for these experiments are presented in Tables 4a–4d. In general, when the classifiers were selected based on the highest F1-score, the AUROC values decreased slightly. However, more balanced PPV and sensitivity can be observed, due to the F1-score being the geometric mean of these.

In mortality classification (Table 4a), an example of the balanced sensitivity and PPV can be seen in the GP-RBF classifier, as sensitivity is increased from 0.157 to 0.845, with the cost of PPV decreasing from 0.493 to 0.253. Similar trade-off can also be seen in the SVM. The RF retains the highest performance in also this experiment with 0.915 AUROC and 0.493 F1-score, however, only a small margin in AUROC in comparison to the Gaussian process classifiers with 0.913 AUROCs. Graphical illustration of the results is presented in Figure 1a.

In BPD classification (Table 4b), the GP-M52 has the highest AUROC of 0.884, however, similar to the mortality classification, the margin is small between the different GP classifiers (0.879) and the RF classifier (0.881). The RF classifier has the highest F1-score with value of 0.704, similar to the results obtained when the AUROC was used for the classifier selection. However, the gap between the RF and the other classifiers has been decreased, as the Gaussian process classifiers and the logistic regression classifiers have around 0.69 F1-scores. Graphical illustration of the results is presented in Figure 1b.

In NEC classification (Table 4c), the RF has the highest AUROC of 0.785 and the highest F1-score of 0.215. We can see that when the Gaussian process classifiers and the SVM classifier were selected based on the highest F1-score, they no longer predict every example as negative, but have generally high sensitivities, the SVM having the highest sensitivity,

**TABLE 4. Classification results for representative classifier selected based on F1-score. The results are presented as the mean and in parentheses the standard error of each evaluation measure, over the cross-validation repetitions. Descriptors under the name of the method are in format: length of time series (hours) - irregularly or regularly sampled (I/R) - all data or exclude first 6 hours (A/E) - class distribution empirical or resampled to uniform (E/U). GP denotes the Gaussian process classifier, with RBF denoting the squared exponential kernel, M32 the Matérn kernel with $\nu = 3/2$, and M52 the Matérn kernel with $\nu = 5/2$. RF denotes the random forest classifier, KNN the $k$-nearest neighbor classifier, LR the logistic regression classifier, LDA the linear discriminant analysis classifier, QDA the quadratic discriminant analysis classifier, and SVM the support vector machine classifier.**

| Method | AUROC | F1-Score | Mean F1-Score | PPV | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| GP-RBF 72h-R-A-U | 0.913 (0.01) | 0.387 (0.02) | 0.641 (0.01) | 0.253 (0.02) | 0.845 (0.04) | 0.821 (0.01) | 0.822 (0.01) |
| GP-M32 72h-R-A-U | 0.913 (0.01) | 0.388 (0.02) | 0.641 (0.01) | 0.253 (0.02) | 0.859 (0.04) | 0.817 (0.01) | 0.819 (0.01) |
| GP-M52 72h-R-A-U | 0.913 (0.01) | 0.387 (0.02) | 0.641 (0.01) | 0.253 (0.02) | 0.855 (0.04) | 0.818 (0.02) | 0.820 (0.01) |
| RF 72h-R-E-E | **0.915** (0.01) | **0.493** (0.03) | **0.720** (0.02) | **0.384** (0.03) | 0.712 (0.05) | 0.918 (0.01) | **0.905** (0.01) |
| KNN 36h-I-E-E | 0.895 (0.03) | 0.437 (0.03) | 0.678 (0.02) | 0.297 (0.02) | 0.839 (0.06) | 0.862 (0.01) | 0.860 (0.01) |
| LR 24h-R-E-E | 0.902 (0.02) | 0.408 (0.02) | 0.657 (0.02) | 0.273 (0.02) | 0.840 (0.05) | 0.837 (0.01) | 0.838 (0.01) |
| LDA 72h-R-A-E | 0.904 (0.01) | 0.390 (0.02) | 0.645 (0.01) | 0.258 (0.02) | 0.828 (0.04) | 0.829 (0.01) | 0.829 (0.01) |
| QDA 72h-R-E-E | 0.832 (0.02) | 0.370 (0.04) | 0.645 (0.02) | 0.265 (0.03) | 0.633 (0.06) | 0.874 (0.01) | 0.859 (0.01) |
| SVM 72h-R-E-U | 0.897 (0.02) | 0.362 (0.02) | 0.622 (0.01) | 0.230 (0.01) | 0.863 (0.04) | 0.797 (0.01) | 0.801 (0.01) |

(a) Mortality

| Method | AUROC | F1-Score | Mean F1-Score | PPV | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| GP-RBF 72h-I-E-U | 0.879 (0.01) | 0.691 (0.01) | 0.762 (0.01) | 0.582 (0.02) | **0.856** (0.02) | 0.754 (0.02) | 0.783 (0.01) |
| GP-M32 72h-I-E-U | 0.879 (0.01) | 0.690 (0.01) | 0.761 (0.01) | 0.581 (0.02) | 0.854 (0.02) | 0.754 (0.02) | 0.782 (0.01) |
| GP-M52 72h-R-A-U | **0.884** (0.01) | 0.690 (0.01) | 0.761 (0.01) | 0.582 (0.01) | 0.851 (0.02) | 0.755 (0.02) | 0.782 (0.01) |
| RF 72h-I-E-E | 0.881 (0.01) | **0.704** (0.02) | **0.785** (0.01) | **0.650** (0.02) | 0.773 (0.03) | **0.833** (0.02) | **0.816** (0.01) |
| KNN 72h-I-A-E | 0.856 (0.01) | 0.683 (0.02) | 0.762 (0.01) | 0.593 (0.02) | 0.810 (0.02) | 0.780 (0.02) | 0.788 (0.01) |
| LR 48h-I-E-E | 0.854 (0.01) | 0.689 (0.01) | 0.767 (0.01) | 0.604 (0.02) | 0.807 (0.02) | 0.789 (0.02) | 0.794 (0.01) |
| LDA 72h-I-A-U | 0.855 (0.01) | 0.684 (0.02) | 0.762 (0.01) | 0.594 (0.02) | 0.813 (0.02) | 0.779 (0.02) | 0.789 (0.01) |
| QDA 72h-R-E-E | 0.842 (0.01) | 0.660 (0.02) | 0.749 (0.01) | 0.595 (0.02) | 0.747 (0.03) | 0.794 (0.02) | 0.781 (0.01) |
| SVM 72h-R-E-U | 0.855 (0.01) | 0.684 (0.02) | 0.763 (0.01) | 0.600 (0.02) | 0.802 (0.03) | 0.785 (0.02) | 0.790 (0.01) |

(b) BPD

| Method | AUROC | F1-Score | Mean F1-Score | PPV | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| GP-RBF 72h-I-A-U | 0.777 (0.04) | 0.126 (0.02) | 0.468 (0.01) | 0.069 (0.01) | 0.702 (0.09) | 0.688 (0.02) | 0.688 (0.02) |
| GP-M32 72h-R-A-U | 0.775 (0.04) | 0.125 (0.01) | 0.465 (0.01) | 0.069 (0.01) | 0.710 (0.08) | 0.682 (0.02) | 0.682 (0.02) |
| GP-M52 72h-I-E-U | 0.764 (0.04) | 0.123 (0.02) | 0.470 (0.01) | 0.068 (0.01) | 0.674 (0.08) | 0.699 (0.02) | 0.698 (0.02) |
| RF 72h-R-A-E | **0.785** (0.03) | **0.215** (0.06) | **0.591** (0.03) | **0.195** (0.07) | 0.271 (0.08) | **0.961** (0.01) | **0.939** (0.01) |
| KNN 72h-I-A-E | 0.718 (0.05) | 0.176 (0.03) | 0.539 (0.02) | 0.105 (0.02) | 0.581 (0.10) | 0.833 (0.01) | 0.825 (0.01) |
| LR 72h-R-A-E | 0.767 (0.04) | 0.145 (0.02) | 0.498 (0.01) | 0.082 (0.01) | 0.678 (0.08) | 0.750 (0.02) | 0.747 (0.01) |
| LDA 36h-R-E-E | 0.759 (0.06) | 0.165 (0.02) | 0.529 (0.01) | 0.097 (0.01) | 0.620 (0.09) | 0.816 (0.02) | 0.810 (0.01) |
| QDA 72h-I-A-E | 0.728 (0.05) | 0.130 (0.03) | 0.529 (0.02) | 0.083 (0.02) | 0.316 (0.09) | 0.885 (0.01) | 0.867 (0.01) |
| SVM 72h-I-E-U | 0.757 (0.04) | 0.121 (0.02) | 0.456 (0.01) | 0.066 (0.01) | **0.741** (0.09) | 0.664 (0.02) | 0.666 (0.02) |

(c) NEC

| Method | AUROC | F1-Score | Mean F1-Score | PPV | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| GP-RBF 72h-I-E-U | 0.843 (0.02) | 0.333 (0.02) | 0.582 (0.01) | 0.209 (0.01) | **0.842** (0.04) | 0.722 (0.01) | 0.731 (0.01) |
| GP-M32 72h-R-E-U | 0.838 (0.01) | 0.332 (0.02) | 0.580 (0.01) | 0.208 (0.01) | 0.840 (0.04) | 0.717 (0.02) | 0.727 (0.01) |
| GP-M52 72h-R-E-U | 0.837 (0.01) | 0.332 (0.02) | 0.580 (0.01) | 0.208 (0.01) | 0.837 (0.04) | 0.719 (0.02) | 0.728 (0.02) |
| RF 72h-R-E-E | **0.846** (0.01) | **0.368** (0.03) | **0.632** (0.02) | **0.261** (0.02) | 0.643 (0.05) | **0.838** (0.01) | **0.822** (0.01) |
| KNN 72h-R-A-E | 0.793 (0.02) | 0.326 (0.02) | 0.583 (0.01) | 0.207 (0.01) | 0.778 (0.04) | 0.739 (0.01) | 0.742 (0.01) |
| LR 72h-R-E-E | 0.837 (0.01) | 0.336 (0.02) | 0.593 (0.01) | 0.217 (0.01) | 0.763 (0.04) | 0.756 (0.02) | 0.757 (0.01) |
| LDA 72h-R-E-E | 0.835 (0.01) | 0.330 (0.02) | 0.586 (0.01) | 0.211 (0.01) | 0.778 (0.04) | 0.742 (0.02) | 0.745 (0.02) |
| QDA 72h-I-E-E | 0.778 (0.02) | 0.315 (0.02) | 0.588 (0.01) | 0.208 (0.01) | 0.669 (0.05) | 0.777 (0.01) | 0.769 (0.01) |
| SVM 72h-I-E-U | 0.826 (0.02) | 0.321 (0.02) | 0.579 (0.01) | 0.203 (0.01) | 0.775 (0.05) | 0.735 (0.01) | 0.739 (0.01) |

(d) ROP

albeit with fairly low PPV. These high sensitivity Gaussian processes and SVM were trained using the class sub-sampling method, which achieves uniform class distribution. We can also see that when these classifiers were selected based on the AUROC, the selected classifiers were trained using empirical distribution. This shows that the type of preprocessing applied to the classifier can improve sensitivity significantly, which may be of special interest in clinical environment. Graphical illustration of the results is presented in Figure 1c.

In ROP classification (Table 4d), the same RF classifier which was the best in AUROC (0.846), also had the highest F1-score (0.368). The margin in F1-score is decreased in relation to the other classifiers, however, it still is relatively high. Similar large improvement in sensitivity can be seen in this task for the Gaussian process classifiers and for the SVM classifier as is seen in the NEC classification. Indeed, the GP-RBF has the highest sensitivity in this task. These high sensitivity classifiers were, again, trained using the class sub-sampling method. Graphical illustration of the results is presented in Figure 1d.

### B. IMPACT OF TIME SERIES LENGTH

In this section, we evaluate the performance of the classifiers with respect to the length of the time series. Figures 2 and 3 visualize model performances, evaluated by AUROC and F1-score, and averaged over the cross-validation repetitions and class-related sampling methods.

From the illustrations, we can see that the exclusion of the first 6 hours of data does not significantly improve or degrade the performance of the classifiers. This indicates that the mean and the standard deviation of the time series can be used for prediction of mortality, BPD, NEC, and ROP without exclusion of the initial measurements.

We can see that in mortality (Fig. 2a), BPD (Fig. 2b), and ROP (Fig. 3b) detection most classifiers have only small benefits from increasing the length of the time series. However, in NEC detection all classifiers seem to benefit from increasing the length of the time series, most noticeably in terms of AUROC. Especially the QDA improves by approximately 0.15 in AUROC. The F1-score of the RF classifier is significantly improved by approximately 0.10 when a longer time series is used.

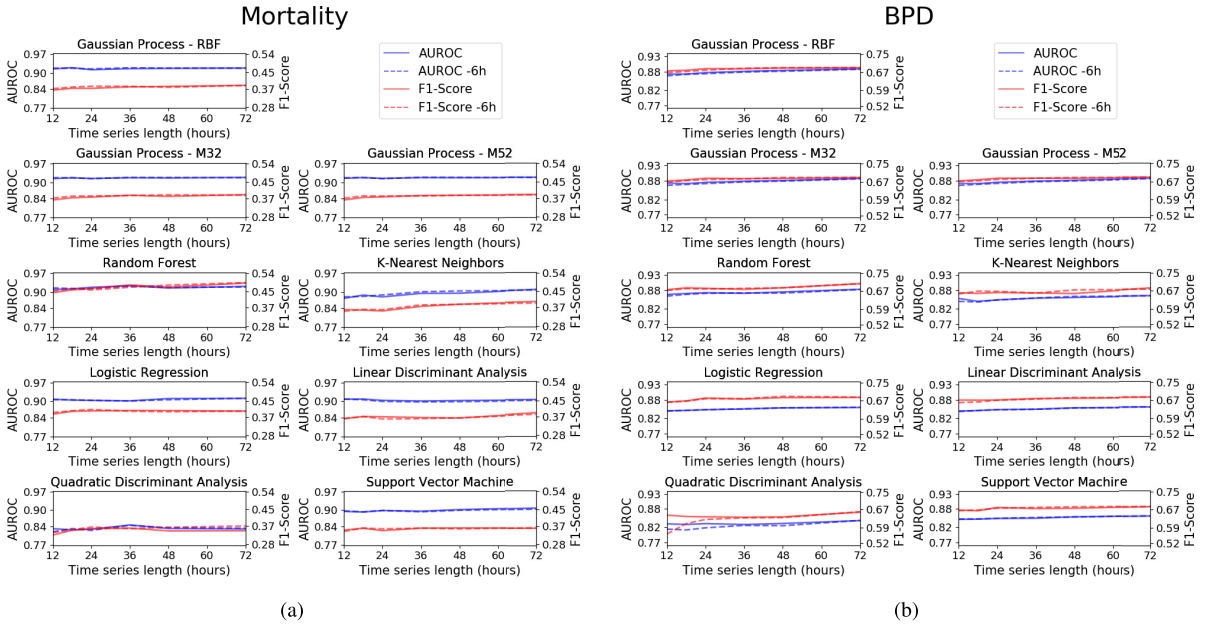



**FIGURE 2.** Effect of time series length on the classification performance for (a) Mortality and (b) BPD. The postfix "−6h" in the legend denotes that the first 6 hours have been removed from the time series.

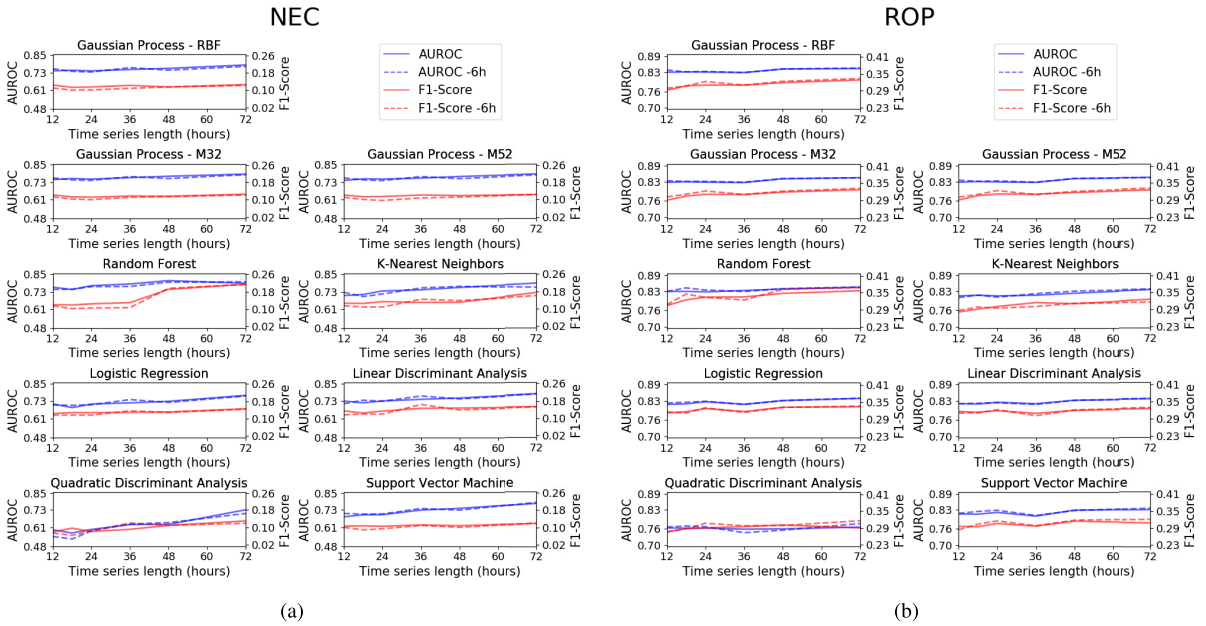



**FIGURE 3.** Effect of time series length on the classification performance for (a) NEC and (b) ROP. The postfix "−6h" in the legend denotes that the first 6 hours have been removed from the time series.

Our analysis suggests that in the prediction of mortality, BPD, and ROP, even 12 hours of time series data could be used for accurate classification, providing earlier detection of these outcomes. In NEC prediction, most classifiers benefit from using longer time series in terms of AUROC, and some also in terms of F1-score, most notably the RF classifier.

### C. FEATURE IMPORTANCE

As the final stage, we used the RF classifier to rank the importance of all extracted features in order to have a better insight to the crucial factors influencing the RF classifier in the detection of neonatal mortality and morbidity. It should be noted that since the RF classifier performs the classification in a nonlinear manner, it is hard to determine the mechanism of how the change in one feature affects the detection performance. However, the feature importances still provide useful information on which features the RF classifier found to be the most informative.

The RF classifier consists of many individual decision trees, each trained using a subset of the patients and features.

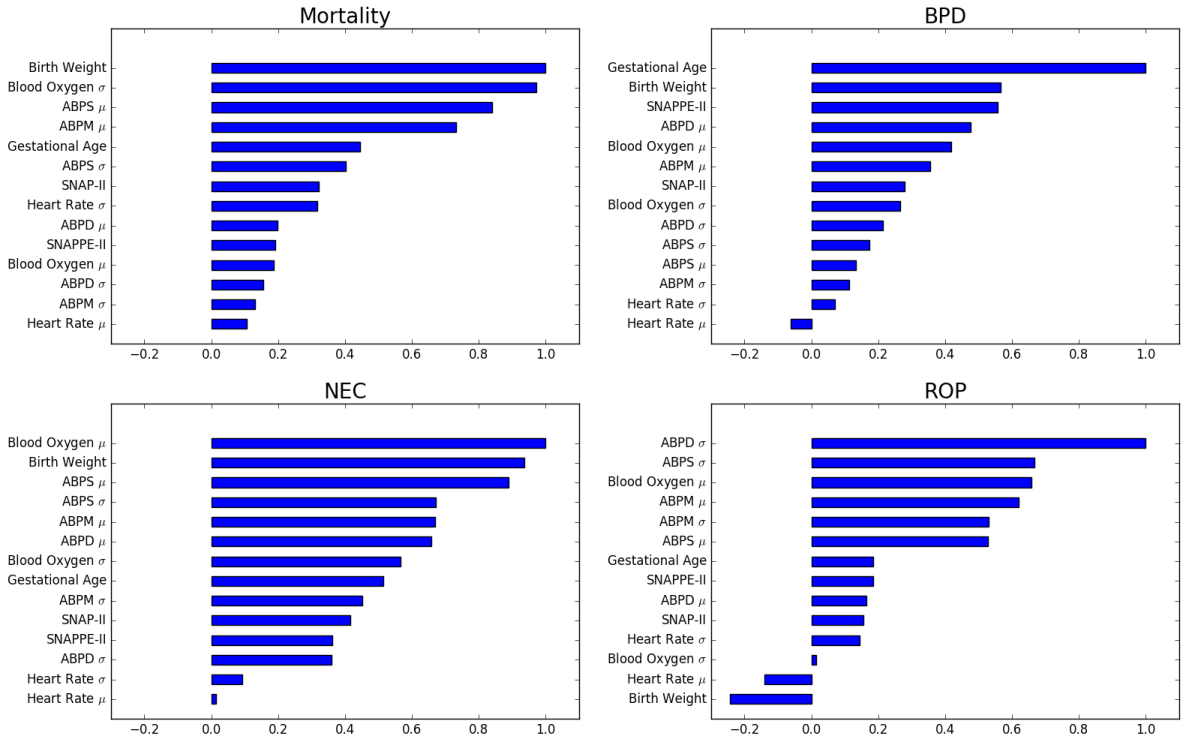



**FIGURE 4.** Feature importances in each task, given by the RF classifier. Here $\mu$ denotes mean, $\sigma$ standard deviation, and ABP denotes arterial blood pressure with postfix D, M and S correspond to diastolic, mean and systolic, respectively. Larger values suggest higher importance. These values are normalized into maximum of 1.0.

The out-of-bag error means the average error of each tree computed on the subset of patients which were not included in training the tree. The out-of-bag error can also be used for evaluating which features are the most influential for the detection performance. This is done by randomly permuting one feature across the different patients and then calculating the our-of-bag error. The feature with the highest out-of-bag error is judged to be the most important one, as randomizing its value caused the highest error. The features with a small positive or negative importance value can be judged to be not important, as randomizing their values does not change the error much [22].

The classifier was trained for each task using all the available data and the longest time series of 72 hours. Feature importances are graphically presented in Figure 4. The importance values have been normalized by the maximum value of all importances. In the case of mortality, birth weight has the largest importance out of all the features, an unsurprising result given the well-known contribution of low birth weight to neonatal mortality [32]. Variation in the blood oxygen saturation has nearly as high importance as birth weight. Mean of systolic and mean arterial blood pressure are the third and fourth most important features. Other features, starting from gestational age, have a large drop in the importance values.

In BPD classification, gestational age has the largest importance by a wide margin. Birth weight and the medical score SNAPPE-II come the second. After the fourth feature-mean of diastolic blood pressure-the feature importances decrease in a linear manner. Low birth weight and gestational age have been previously discovered to contribute to increased risk of BPD [33]. Also, increased SNAPPE-II has been associated with BPD [34]. Mechanical ventilation and supplementary oxygen are thought to play a major factor in the development of BPD [33], which might explain the blood oxygen concentration being in the top half of the features.

In NEC classification, blood oxygen concentration, birth weight and the mean of systolic arterial blood pressure were the three most important features, after which a large drop in the numerical value of importances is seen. In ROP classification, the standard deviation of diastolic arterial blood pressure had the largest importance, by a relatively wide margin. The following features in descending order were the standard deviation of systolic blood pressure, the mean of blood oxygen, the mean of mean arterial blood pressure, the standard deviation of mean arterial blood pressure, and the mean of systolic blood pressure. However, they had similar importances. After these features, a large drop in importance is seen.

Interesting observation can be seen when comparing mortality, BPD, and NEC feature importances with the ones of ROP. Within the first group, birth weight is either the most or the second most important feature, however, it is the least important feature in ROP classification. The most important feature in ROP classification-the standard deviation of diastolic blood pressure-is also the third least important feature in

mortality and NEC classification, and the sixth least important feature in BPD classification.

## VII. CONCLUSIONS

In this article, we have presented a systematic analysis of 9 different classifiers in the tasks of neonatal mortality, BPD, NEC, and ROP detection, using an NICU dataset. Our preprocessed datasets included irregular and regularized time series, time series of different lengths, and time series with initial 6 hours of observations excluded. Our experiments also included majority class sub-sampling for uniform class distribution. In our experiments the RF classifier had generally the best performance, when evaluated using our primary evaluation measures of AUROC and F1-score.

Our results show that the features proposed in the preliminary work [13], [14] are robust to the choice of classifiers in AUROC and F1-score measures, when the training data has been preprocessed in an optimal fashion. Our findings also show that the performance of the classifiers is unaffected by the exclusion of the first 6 hours of data. This finding indicates that the possible heteroscedasticity present in the neonatal adaptation period measurements does not degrade the performance of the classifiers.

In comparison to our preliminary work, presented in [14] and [13], we have shown that our preprocessing methods and the proposed F1-score for classifier selection improve the sensitivity of the same classifiers in the task of neonatal mortality prediction, albeit this methodology decreased the AUROC values slightly. In the task of BPD, NEC, and ROP classification, we have shown that the RF classifier can reach the best or the second best results in AUROC, and the best results in F1-score in each task. However, other classifiers also reach competitive F1-scores when the preprocessing scheme is tuned to maximize F1-score.

Our results suggest that in a clinical setting, where the sensitivity can have a high priority, the AUROC evaluation measure might not provide a good standard for comparison. Instead, the F1-score could provide a useful measure for comparison. The overall results also show that when sensitivity is important, sub-sampling the majority class can lead to increased sensitivity. In majority of the cases and in majority of the classifiers, the best results were obtained in each primary evaluation measure when the longest 72 hour period of data was used. This suggests that the long term mean and standard deviation of the time series provide better discriminative features than the short term counterparts. However, this effect was not very significant in other tasks than NEC classification.

Lastly, we acknowledge some limitations of this study, namely i) the dataset was collected in the same hospital, possibly introducing challenges to generalization to other hospital settings with different devices for physiological signal measurements, and ii) our algorithm is designed in a retrospective manner after the measurements for a certain time period are taken. This requires some adaptations for real-time applications, which possibly leads to a drop in the performance.

## REFERENCES


[1] M. S. Harrison and R. L. Goldenberg, "Global burden of prematurity," *Seminars Fetal Neonatal Med.*, vol. 21, no. 2, pp. 74–79, Apr. 2016.

[2] R. J. Baer, E. E. Rogers, J. C. Partridge, J. G. Anderson, M. Morris, M. Kuppermann, L. S. Franck, L. Rand, and L. L. Jelliffe-Pawlowski, "Population-based risks of mortality and preterm morbidity by gestational age and birth weight," *J. Perinatol.*, vol. 36, no. 11, pp. 1008–1013, Nov. 2016.

[3] J. D. Horbar, E. M. Edwards, L. T. Greenberg, K. A. Morrow, R. F. Soll, M. E. Buus-Frank, and J. S. Buzas, "Variation in performance of neonatal intensive care units in the united states," *JAMA Pediatrics*, vol. 171, no. 3, Mar. 2017, Art. no. e164396.

[4] *Levels & Trends in Child Mortality: Report 2018, Estimates Developed by the United Nations Inter-Agency Group for Child Mortality Estimation*, United Nations ChildrenÃ Fund, New York, NY, USA, 2018.

[5] L. Davidson and S. Berkelhamer, "Bronchopulmonary dysplasia: Chronic lung disease of infancy and long-term pulmonary outcomes," *J. Clin. Med.*, vol. 6, no. 1, p. 4, Jan. 2017.

[6] A. M. Thompson and M. J. Bizzarro, "Necrotizing enterocolitis in newborns," *Drugs*, vol. 68, no. 9, pp. 1227–1238, 2008.

[7] A. Hellström, L. E. Smith, and O. Dammann, "Retinopathy of prematurity," *Lancet*, vol. 382, pp. 1445–1457, Oct. 2013.

[8] M. A. Isani, P. T. Delaplain, A. Grishin, and H. R. Ford, "Evolving understanding of neonatal necrotizing enterocolitis," *Current Opinion Pediatrics*, vol. 30, no. 3, pp. 417–423, Jun. 2018.

[9] V. Apgar, "A proposal for a new method of evaluation of the newborn Infant.," *Anesthesia Analgesia*, vol. 32, no. 1, pp. 260–267, Jan. 1953.

[10] D. K. Richardson, J. E. Gray, M. C. McCormick, K. Workman, and D. A. Goldmann, "Score for neonatal acute physiology: A physiologic severity index for neonatal intensive care," *Pediatrics*, vol. 91, no. 3, pp. 617–623, 1993.

[11] D. K. Richardson, C. S. Phibbs, J. E. Gray, M. C. McCormick, K. Workman-Daniels, and D. A. Goldmann, "Birth weight and illness severity: Independent predictors of neonatal mortality," *Pediatrics*, vol. 91, no. 5, pp. 969–975, 1993.

[12] D. K. Richardson, J. D. Corcoran, G. J. Escobar, and S. K. Lee, "SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores," *J. Pediatrics*, vol. 138, no. 1, pp. 92–100, Jan. 2001.

[13] O.-P. Rinta-Koski, S. Särkkä, J. Hollmén, M. Leskinen, and S. Andersson, "," Gaussian process classification for prediction of in-hospital mortality among preterm infants," *Neurocomputing*, vol. 298, pp. 134–141, Jul. 2018.

[14] O.-P. Rinta-Koski, S. Sarkka, J. Hollmen, M. Leskinen, K. Rantakari, and S. Andersson, "Prediction of major complications affecting very low birth weight infants," in *Proc. IEEE Life Sci. Conf. (LSC)*, Dec. 2017, pp. 186–189.

[15] E. Baraldi and M. Filippone, "Chronic lung disease after premature birth," *New England J. Med.*, vol. 357, no. 19, pp. 1946–1955, Nov. 2007.

[16] J. Neu and W. A. Walker, "Necrotizing enterocolitis," *New England J. Med.*, vol. 364, no. 3, pp. 255–264, 2011.

[17] B. W. Fleck and N. McIntosh, "Pathogenesis of retinopathy of prematurity and possible preventive strategies," *Early Human Develop.*, vol. 84, no. 2, pp. 83–88, Feb. 2008.

[18] H. A. Mintz-Hittner, K. A. Kennedy, and A. Z. Chuang, "Efficacy of intravitreal bevacizumab for stage 3+ retinopathy of prematurity," *New England J. Med.*, vol. 364, no. 7, pp. 603–615, Feb. 2011.

[19] B. M. Casey, D. D. McIntire, and K. J. Leveno, "The continuing value of the apgar score for the assessment of newborn infants," *New England J. Med.*, vol. 344, no. 7, pp. 467–471, Feb. 2001.

[20] S. Saria, A. K. Rajani, J. Gould, D. Koller, and A. A. Penn, "Integration of early physiological responses predicts later illness severity in preterm infants," *Sci. Transl. Med.*, vol. 2, no. 48, p. 48ra65–48ra65, 2010.

[21] M. Podda, D. Bacciu, A. Micheli, R. Bellá, G. Placidi, and L. Gagliardi, "A machine learning approach to estimating preterm infants survival: Development of the preterm infants survival assessment (PISA) predictor," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 13743.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements Stat. Learning: Data Mining, Inference, Prediction*, 2nd Ed. New York, NY, USA: Springer, 2009.

[23] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer-Verlag, 2006.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[25] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
[26] C. K. I. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.
[27] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997.
[28] V. Gulshan, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.
[29] D. S. W. Ting, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, pp. 2211–2223, 2017.
[30] J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari, "GPstuff: Bayesian modeling with Gaussian processes," *J. Mach. Learn. Res.*, vol. 14, no. Apr, pp. 1175–1179, 2013.
[31] F. Pedregosa, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[32] M. C. McCormick, "The contribution of low birth weight to infant mortality and childhood morbidity," *New England J. Med.*, vol. 312, no. 2, pp. 82–90, Jan. 1985.
[33] E. Bancalari, N. Claure, and I. R. S. Sosenko, "Bronchopulmonary dysplasia: Changes in pathogenesis, epidemiology and definition," *Seminars Neonatol.*, vol. 8, no. 1, pp. 63–71, Feb. 2003.
[34] B. Özcan, A. S. Kavurt, O. Aydemir, Z. Gençtürk, A. Y. Baş, and N. Demirel, "SNAPPE-II and risk of neonatal morbidities in very low birth weight preterm infants," *The Turkish J.*, vol. 59, no. 2, pp. 105–112, 2017.

**JOEL JASKARI** received the B.Sc. (Tech.) and M.Sc. (Tech.) degrees from Aalto University, Finland, in 2013 and 2017, respectively, where he is currently pursuing the D.Sc. degree with the Department of Computer Science, School of Science. His research interests include machine learning, especially deep learning and Bayesian methods, and the application of machine learning in healthcare.

**JANNE MYLLÄRINEN** received the B.Sc. (Econ.) and M.Sc. (Tech.) degrees from Aalto University, Finland, in 2016 and 2019, respectively. His main research interests include emerging data science related technologies, such as artificial intelligence and machine learning, and their applications in health, energy, and finance.

**MARKUS LESKINEN** is currently a Neonatology Consultant with the Childrens' Hospital, Helsinki, which was one of the early adapters of electronic patient Information systems in NICU with comprehensive electronic archives running back to 1999. His current research interests include machine learning in neonatology and data driven quality improvement in collaboration with the School of Science, Aalto University. He has authored or coauthored 15 peer-reviewed scientific articles.

**ALI BAHRAMI RAD** (Member, IEEE) received the B.Sc. degree in electrical engineering from Imam Hossein University, Tehran, Iran, in 2003, the M.Sc. degree in biomedical engineering from the Tampere University of Technology, Tampere, Finland, in 2011, and the Ph.D. degree in computer engineering-information technology from the University of Stavanger, Stavanger, Norway, in 2017. He completed his Postdoctoral Research with the Department of Electrical Engineering and Automation, Aalto University, Espoo, Finland. He is currently a Visiting Assistant Professor with the Department of Biomedical Informatics, Emory University, Atlanta, USA. His current research interests include machine learning, neural systems, and biomedical/electrophysiological signal analysis with applications to resuscitation, defibrillation, seizure detection, and stem-cell technology. He was a recipient of multiple international prizes in biosignal analysis and machine learning contests, including the PhysioNet/Computing in Cardiology Challenge 2017 (tied for the 1st place), the PhysioNet/Computing in Cardiology Challenge 2016 (2nd place), and the Brain-Computer Interface Challenge at IEEE EMBS Neural Engineering Conference 2015 (3rd place).

**JAAKKO HOLLMÉN** (Senior Member, IEEE) received the D.Sc. (Tech.) degree from the Helsinki University of Technology, in 2000. He has held various positions at the Helsinki University of Technology and Aalto University, Finland, since then. In 2019, he joined the Data Science Research Group, Stockholm University. He is currently an Associate Professor with the Department of Computer and Systems Sciences, Stockholm University, Sweden. His research interests include machine learning and data mining, with applications in health and medicine as well as environmental applications in the analysis of data in built and natural environments. He has published over 140 scientific publications in these fields. He has organized and led many international conferences in these fields. He is an Editor of the *Intelligent Data Analysis* journal and a member of the Editorial Board of *Data Mining and Knowledge Discovery*.

**STURE ANDERSSON** received the M.D. and Ph.D. degrees. He is currently an Emeritus Professor of neonatology, and one of the pioneers in using big data in neonatology. He has authored or coauthored over 300 peer-reviewed scientific articles.

**SIMO SÄRKKÄ** (Senior Member, IEEE) is currently an Associate Professor with Aalto University. He has authored or coauthored over 100 peer-reviewed scientific articles and three books. His research interests include multi-sensor data processing systems and machine learning methods with applications in medical and health technology, target tracking, inverse problems, and location sensing. He is a member of the IEEE Machine Learning for Signal Processing Technical Committee. He has been serving as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS.